Structural characterization of gene products from uncharacterized operons of *Pseudomonas aeruginosa* PAO1

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades

eines Doktors der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Christian Gerhard Feiler
aus Oberviechtach

1. Referent: Prof. Dr. Wulf Blankenfeldt
2. Referent: Prof. Dr. Michael Steinert
eingereicht am: 11.02.2015
mündliche Prüfung (Disputation) am: 17.07.2015

Druckjahr 2015

Table of content

1. Introduction 8
   1.1. Bacteria 8
   1.2. *Pseudomonas aeruginosa* 9
   1.3. Annotation problem 11
   1.4. A novel approach 16
2. Aim of this study 18
3. Materials 19
   3.1. Chemicals 19
   3.2. Enzymes and Kits 19
   3.3. Bacterial Strains 20
   3.4. Media 20
   3.5. Antibiotics 21
   3.6. Buffer and solutions 21
   3.7. Bacterial strains, plasmids and primer 22
4. Methods 23
   4.1. PCR and molecular cloning 23
   4.1.1. Preparation of genomic DNA 23
   4.1.2. Extraction of plasmid DNA 23
   4.1.3. Polymerase chain reaction (PCR) 24
   4.1.4. DNA restriction and cloning 25
   4.1.5. Colony PCR and Sequencing 26
   4.1.6. QuikChange mutagenesis 27
   4.2. Bacterial transformation - general aspects 27
   4.2.1. Electro competent cells 28
   4.2.2. Chemical competent cells 28
   4.2.3. Transformation of genetic material 29
   4.2.4. Bacterial storage 29
   4.3. Expression plasmids – general aspects 29
   4.3.1. Generation of *p5$* 30
   4.3.2. Engineering of *p10$* 31
   4.3.3. Construction of *pCryst* and *pUrgé* 32

4.3.4. Plasmids constructed and used in this work 32
4.4. Gene expression and protein purification 33
4.4.1. General aspects and plasmid design 33
4.4.2. Native protein expression 35
4.4.3. Expression of selenomethionine labeled protein 35
4.5. Protein purification 37
4.5.1. Metal affinity chromatography – general aspect 37
4.5.2. Nickel affinity chromatography 37
4.5.3. Size exclusion chromatography 38
4.5.4. TEV protease 40
4.5.5. Rhino-Virus 3C Protease 42
4.5.6. SUMO protease 43
4.5.7. Purification of PncA - Nicotinamidase 43
4.6. Analytical methods 44
4.6.1. Agarose gel electrophoresis 44
4.6.2. SDS-Polyacrylamid gel-electrophoresis 44
4.6.3. Determination of DNA and protein concentration 45
4.6.4. PA5507 enzyme activity test 45
4.6.5. SAXS data collection 46
4.7. Genetic manipulation of *Pseudomonas aeruginosa PAO1* 46
4.7.1. Operon selection 46
4.7.2. Homologous recombination and operon knock-out 47
4.7.3. Extraction protocol for metabolomic analysis 49
4.8. Principles of Crystallography 51
4.8.1. Concept 51
4.8.2. Protein crystals 52
4.8.3. Bragg's Law 54
4.8.4. Real space and reciprocal lattice 56
4.8.5. Bragg's law in reciprocal space – Ewald construction 58
4.8.6. Diffraction to electron density – Fourier transformation 59
4.8.7. Molecular replacement and experimental phasing 63
4.8.8. Physical background – anomalous scattering 65
4.8.9. Model building, refinement and structure validation 66
4.9. X-ray crystallography methods 69
4.9.1. Crystallization 69

4.9.2. Data collection 69
4.9.3. Data processing 71
4.9.4. Structure and phase determination 72
5. Results and Discussion: 74
5.1. Structural analysis of PA1622 74
5.1.1. Sequence analysis of PA1622 74
5.1.2. Structural prediction 76
5.1.3. Structure of PA1622 78
5.1.4. Domain organization 83
5.1.5. Structural comparison to similar predicted structures 86
5.1.6. Active site of PA1622 87
5.1.7. Homologs regulate activity involving lid domain movement 90
5.1.8. Structure in context 91
5.2. Structural analysis of PA1623 94
5.2.1. Sequence analysis 94
5.2.2. Note of status 96
5.2.3. Short background to Glutathione-S-transferase 96
5.2.4. Structure solution of PA1623 96
5.2.5. Structure of apo PA1623 99
5.2.6. Structure of PA1623 bound to glutathione 102
5.2.7. The C-terminus generates an access tunnel to GSSG 106
5.2.8. Structure of PA1623 in context 108
5.3. Structural analysis of PA1624 112
5.3.1. Sequence analysis of PA1624 112
5.3.2. PA1624 - unpredictable protein fold 112
5.3.3. Structure solution of PA1624 114
5.3.4. PA1624 Structure and homology search 117
5.4. Structural analysis of PA5506 122
5.4.1. Sequence analysis of PA5506 122
5.4.2. PA5506 – structural prediction 124
5.4.3. Purification, data collection and structure solution of PA5506 125
5.4.4. PA5506 crystallized in truncated form 130
5.4.5. Structure of PA5506 133
5.4.6. PA5506 has flexible termini 136
5.4.7. Metal binding involves a C-terminal rearrangement 137

5.4.8. Ligand binding site identification 140
5.4.9. Structure in context 142
5.5. Structural analysis of PA5507 144
5.5.1. PA5507 sequence analysis 144
5.5.2. Structure solution and model building 145
5.5.3. Structure of PA5507 148
5.5.4. Homology and active site analysis 152
5.5.5. Initial activity tests with amide substrates 153
5.5.6. Structure in context 154
6. Summary 157
7. Outlook 160
8. Acknowledgement 161
9. Appendix 163
9.1 Additional list of primers used in this study 163
9.2 Data collection statistics 164
9.3 Structural prediction 166
9.4 Protein expression and purification 167
9.4.1 PA5506 expression and purification 167
9.4.2 PA5507 expression and purification 168
9.4.3 PA5508 expression and purification 168
9.4.4 PA1622 expression and purification 168
9.4.5 PA1623 expression and purification 169
9.4.6 PA1624 expression and purification 169
9.5 Crystallization 170
9.5.1 Screening and optimization 170
9.6 Fluorescense scan 171
9.6.1 SAD phasing of PA5506/PA1622/PA1623/PA1624 171
9.7 Plasmid maps 172
10. References 173
11. Curriculum Vitae 188

Für meine Großeltern

1. Introduction

1.1. Bacteria

Microorganisms are constantly surrounding us in every situation of life. Their appearance on earth was dated to more than 3.5 billion years ago (Altermann & Kazmierczak, 2003; Schopf, 2006) and they have evolved ever since, which renders them the most adapted organisms on the planet. During the 'Great Oxygen Event' (Schirrmeister et al., 2013), photosynthetically active cyanobacteria started to introduce oxygen into the atmosphere and thereby changed the environment (Tice & Lowe, 2004; Konhauser et al., 2011) (figure 1). Even in the absence of sunlight due to dirt and rock hurtled into the atmosphere, bacteria species were able to produce oxygen (Ettwig et al., 2010).

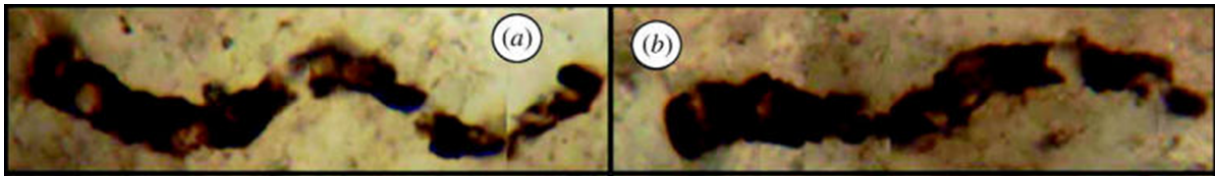


Figure 1: Archaean microfossils with cyanobacterium-like morphology (Schopf, 1993) dated to 3.5 billion year old were identified found in today's Western Australia (Schopf, 2006) (a-b).

Their impossible ability to reproduce sexually made them acquire different mechanisms to adapt to new environmental conditions using horizontal exchange of genetic material (Arber, 2014), which allows them to populate even the most hostile niches. Due to their adaptability, bacteria have also colonized humankind. Most of the bacteria are harmless to men and moreover, humanity is ever since dependent on the existence of bacteria. For example our digestive system contains ten times more adjuvant bacteria then cells in the entire body (Mukherjee & Hooper, 2015). Besides the number of beneficial bacteria, a smaller number are directly pathogenic to human. In contrast to advantageous symbiotic living bacteria, pathogenic bacteria usually grow in a very diverse and at the same time rapidly evolving community. Short generation times and several methods of genome modulation allow them to rapidly adapt to new environmental conditions. Even though bacteria are surrounding us every day, the human immune system is able to protect us from most pathogenic bacteria by several stages of host defense mechanism. In some cases an external antimicrobial therapy is needed to fight the pathogen. However, with their ability to adapt to antibiotical treatments, some pathogens became resistant. Some

of the most prominent and emerging pathogens in today's western world are for example *Helicobacter pylori* (Kusters et al., 2006), *Campylobacter* (O'Donovan et al., 2014), *Mycobacterium tuberculosis* (Modlin & Bloom, 2013) and *Pseudomonas aeruginosa* (Sousa & Pereira, 2014; Tümmler et al., 2014).

1.2. *Pseudomonas aeruginosa*

Recently the Infectious Diseases Society of America (IDSA - http://www.idsociety.org) identified a clique of bacteria that manage to escape lethal action of antibiotics by developing sophisticated resistant mechanism (Pendleton et al., 2013). Besides *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumonia and Acinetobacter baumanni, Pseudomonas aeruginosa* was named as a member the "*ESKAPE"* pathogens, as well as *Enterobacter species* (Rice, 2008) (figure 2A).

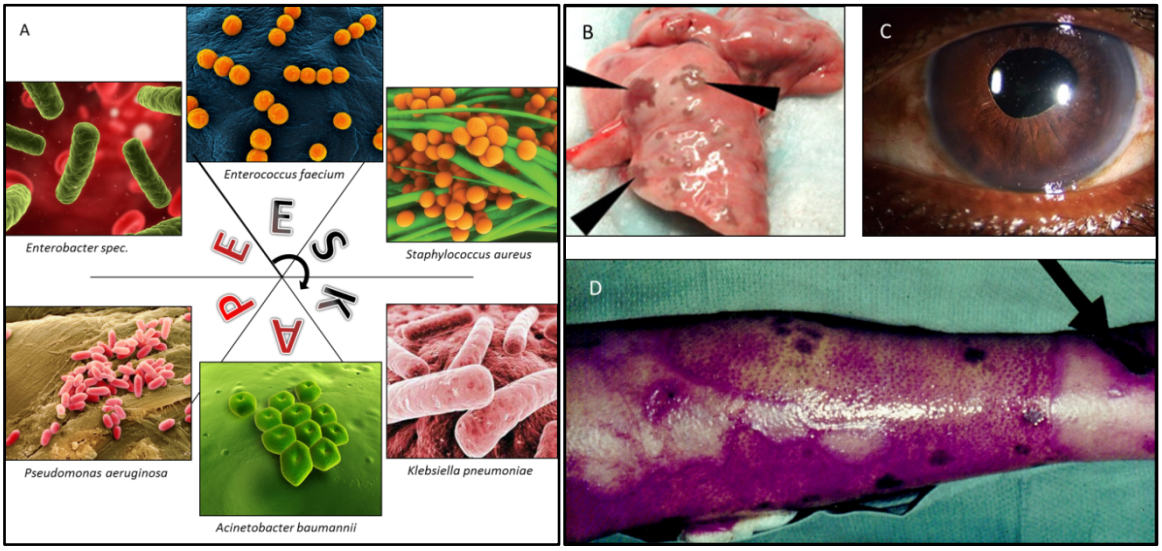


Figure 2: Members of the ESKAPE clique (A), images were adapted from bioquell (ww.bioquell.ie). *Pseudomonas aeruginosa* infections of the lungs (B) (Ochsner et al., 2002), eye (C) (Bharathi et al., 2014) and skin (D) (Pruitt, 2000).

These prevalent opportunistic human and animal pathogens can feed on a large variety of substrates and shows new paradigms in pathogenesis, transmission and resistance (Peterson, 2009). The gram-negative flagellated bacteria *Pseudomonas aeruginosa* is able to colonize a variety of different host organism ranging from insects and animals to plants and mammals (Mahajan-Miklos et al., 1999; D'Argenio et al., 2001; Walker et al., 2004). It is ubiquitous in nature (Green et al., 1974) and has the ability to colonize adverse environments. It is able to utilize unusual carbon sources like diesel or jet fuel, where it is

known as hydrocarbon utilizing microorganism (HUMbug) causing corrosion of fuel systems and water pipes (Yuan & Pehkonen, 2007; Itah et al., 2009). With their metabolic versatility, *Pseudomonads* utilize compounds that are highly toxic to other bacterial species, such as xylene or toluene (Adair et al., 1969; Williams & Worsey, 1976) or even chemical disinfectant (Lakkis & Fleiszig, 2001). The ability to feed on this wide range of substrates as well as its high temperature tolerance ranging from 4°C up to 42°C provides prominent evolutionary advantage over other bacteria in colonizing previously untouched niches (Tsuji et al., 1982; Tümmler et al., 2014).

Pseudomonas is responsible for a number of infections including meningitis, endocarditis and systemic infections. As *Pseudomonas aeruginosa* also colonizes hospital reservoirs, it is a serious threat to the immune compromised such as AIDS patients or individuals suffering from neutropenia (Bendig et al., 1987; Aloush et al., 2006) and cystic fibrosis (CF). It is one of the most commonly found gram-negative bacilli in nosocomial infections (Klevens et al., 2007; Hidron et al., 2008) responsible for more than 10% of all cases (Aloush et al., 2006). Furthermore, it is the most common pathogenic bacteria isolated from patients that have been hospitalized for more than a week (Lessnau et al., 2014). Additionally, burn victims which partially lack the protective skin layer as well as patients suffering from cystic fibrosis, where *Pseudomonas* colonizes the thick mucus in the lungs, are targeted hosts (Lyczak et al., 2000; Pendleton et al., 2013). Other than medical treatments of infections caused by related gram-negative bacteria, *Pseudomonas aeruginosa* infections are more demanding and severe. Bacterial infection and virulence is usually restricted to a certain part of the host, *Pseudomonas aeruginosa* however is capable of infecting virtually all tissue (Lyczak et al., 2000). Bacterial virulence factors are injected into the host cells via the type III secretion system and interfere with e.g. host's small GTPase cascades (Krall et al., 2000; Sawa, 2014), effect the signaling by cAMP (Yahr et al., 1998) and potentially act as cytotoxin causing cell death (Hauser et al., 1998). Elastase, phospholipase and polysaccharide aliginate are secreted during all stages of infection to protect the pseudomonad from host immune response. This enables the bacteria to undermine host defense while striking back by a mechanism called "pack swarming" whereby the bacteria simultaneously attack and destroy one host cell (Dacheux et al., 2000, 2001).

Bacterial infections have been extensively treated by antibiotic application within the last five decades. Amongst other bacteria, *Pseudomonas* became resistant to many antimicrobials (Roy et al., 2010). Beta-lactams are ineffective (Bassetti et al., 2013) due to the bacteria's natural resistance mediated by multiple extended-spectrum beta-lactamases (*Ambler class A to D*) (Lister et al., 2009; Hakemi Vala et al., 2014). Moreover, *P. aeruginosa* employs a sophisticated set of multi drug efflux pumps (Ozer et al., 2012; Morita et al., 2015) and possesses a modified membrane envelope that enforces an exclusion limit of 500 Da via a membrane porin (Livermore, 2002). Some strains have been reported to be resistant to all common antibiotics also including aminoglycosides, fluoroquinolones, cephalosporins and carbapenems (Vanhoof et al., 1993). Gained antibiotic resistance is distributed rather fast within a certain environmental habitat, also due to elevated gene transfer (Darch et al., 2015).

Recent reports about pan-drug resistant strains of *P. aeuruginosa*, non-susceptible to any antimicrobial agent that were isolated from several clinics affirm the drastic situation (Tan et al., 2014). An infection can result in a life threatening outcome since *Pseudomonas* disseminates by the bloodstream, causing inflammation in any part of the body in later stages of infection, at which an antimicrobial treatment is almost impossible (van Delden, 2007; Shorr, 2009; Lessnau et al., 2014). Therefore, it is one of the most prevalent bacteria in nosocomial infections (Klevens et al., 2007; Hidron et al., 2008; Bereket et al., 2012), the leading pathogenic source of morbidity and mortality in individuals afflicted with cystic fibrosis (Hogardt & Heesemann, 2013) and one of the three top pathogens in patients with sepsis syndrome (Tamme et al., 2000).

1.3. Annotation problem

The first complete sequenced genome was that of *Haemophilus influenza* in 1995, a bacteria causing severe infections in human (Fleischmann et al., 1995). Later, sequences of lower eukaryotic genomes became available (Engel et al., 2013), followed by genomes of multicellular organisms (Consortium, 1998; Adams et al., 2000). The human sequencing project (HGP) marked a peak in sequencing effort with the release of the first draft of a complete human genome (McPherson et al., 2001; Venter et al., 2001). With the transition from classical 'first-generation' Sanger-based sequencing (Sanger et al., 1977)

to ‘next-generation’ sequencing in 2008 (Schuster, 2008; Metzker, 2010), a rapid increase of deposited genome sequences by at the same time significantly lower costs was achieved (figure 3) (2010; Stein, 2010; Mardis, 2011), resulting in the proposal of the ‘$1000 genome’ (Liebert, 2014).

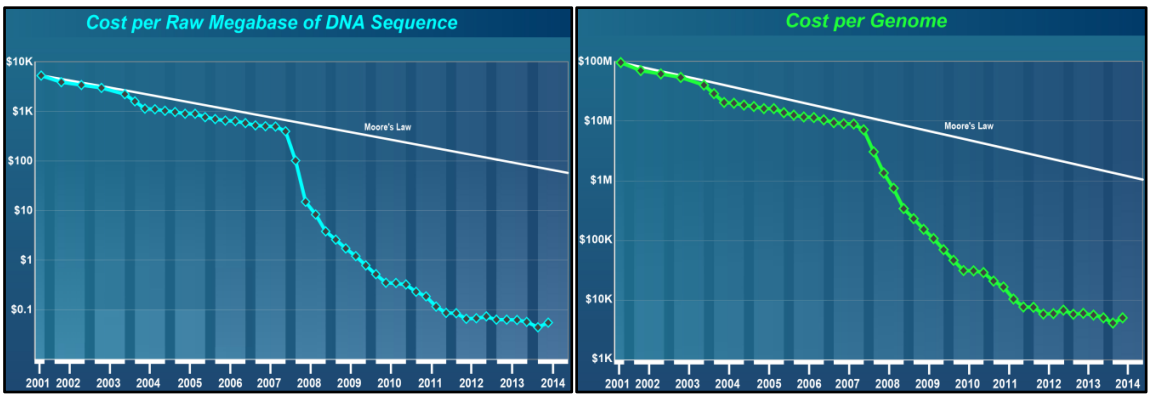


Figure 3: Cost of DNA sequencing per megabase and average sized genome; note the logarithmic scale on the Y axis. Profound out-pacing of Moore's Law in January 2008 is reflected in the transition from Sanger-based to 'next-generation' DNA sequencing technologies; with compliance adopted from NIH (www.genome.gov).

High-throughput sequencing made new whole-genome sequences available on daily basis, with more than 25 genomes published every day in 2013 (Pruitt et al., 2002, 2012, 2014; Tatusova et al., 2014). Since 1995 the exponential growth kept steady, and if the current trend holds, a total of 100,000 genomes sequences will be available between 2017 and 2020 (figure 4).

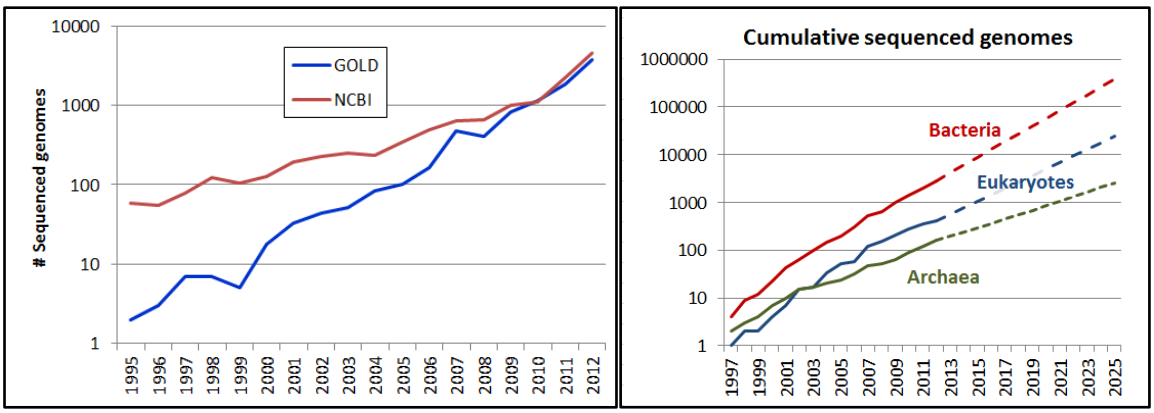


Figure 4: Left: In 2012, GOLD and NCBI added 3736 and 4585 genome sequences to the database (NCBI), respectively. Right: Excerpt from the GOLD database of “completed data” fully sequenced genomes, sorted by year of complete release and split by kingdoms.

Knowledge about the genetic sequence of an organism becomes more and more important in today's life sciences research. However, it alone does not allow scientists to find answers to demanding questions in the world. The breakthrough announced by Bill Clinton shortly before the release of the first draft of the human genome (Venter et al., 2001) that this would "revolutionize the diagnosis, prevention and treatment of most, if not all, human diseases" remained largely elusive (Wade, 2010). Initially, scientists believed that by sequence analysis they would be able to ferret out the roots of genetic diseases including Alzheimer's and certain types of cancer (Evans et al., 2011). This euphorism was drowned when it became clear that knowledge of genetic composition cannot be directly reflected to treatment of disease. This holds true not only for the result of the HGP but is true for all new sequenced organisms. Moreover, newly sequenced genomes, especially those of procaryotic organisms, contains 40-50% of genes flagged as "hypothetical proteins" with unknown function (Schnoes et al., 2009). This also applies to *Pseudomonas aeruginosa,* however, from the time when the bacteria's genome was sequenced (NCBI ref: NC_002516) (Stover et al., 2000) and initial genome annotations were carried out, information and research conclusions were gathered in one online resource database, the *Pseudomonas* Genome Database (www.pseudomonas.com) (Winsor et al., 2011). New sequences and annotation are deposited to NCBI as well as are documented in the *Pseudomonas* Genome Database (Winsor et al., 2011), generating a *Pseudomonas* specific annotation database.

Recent genome sequencing projects revealed previously inconceivable details about the *Pseudomonas aeruginosa* genome. The broad metabolic capacity and ubiquitous distribution originates from the genetic repertoire which is larger than the human genome (Tümmler et al., 2014). Common amongst all *Pseudomonas* species are about 4000 genes, the core genome (Römling et al., 1995; Spencer et al., 2003). More than 10.000 accessory genes are found in several strains that complement the genome with so-called regions of genome plasticity (Mathee et al., 2008; Klockgether et al., 2011; Stewart et al., 2011; Valadbeigi et al., 2014) resulting in genome size variations between 5.5 and 7 Mbps (Schmidt et al., 1996; Lee et al., 2006).

The genome of *Pseudomonas aeruginosa* PAO1, the most commonly used strain for research, covers 5678 open reading frames (ORFs), clustered into ~1200 operons. Operons are genes that are transcribed to polycistronic mRNA, that translate to several

proteins and that are usually functionally linked. The *P. aeruginosa* PAO1 ORFs are clustered in 1139 operons of variable size, which can be predicted by bioinformatic tools with high precision (Mao et al., 2014).

Amongst all the ORFs, only seventeen percent are experimentally characterized in *Pseudomonas aeruginosa* PAO1. Furthermore, the function of about 45% of all genes is annotated according to experimental demonstration in closely homologous organisms or sequence homology. However, closer inspection reveals that about 2300 ORFs, 40% of the genome, still remain functionally uncharacterized with homology to previously reported genes of unknown function, or no similarity to any previously reported sequence (figure 5). Considering this number, it is clear that new discoveries are most likely hidden inside this uncharacterized genome fraction.

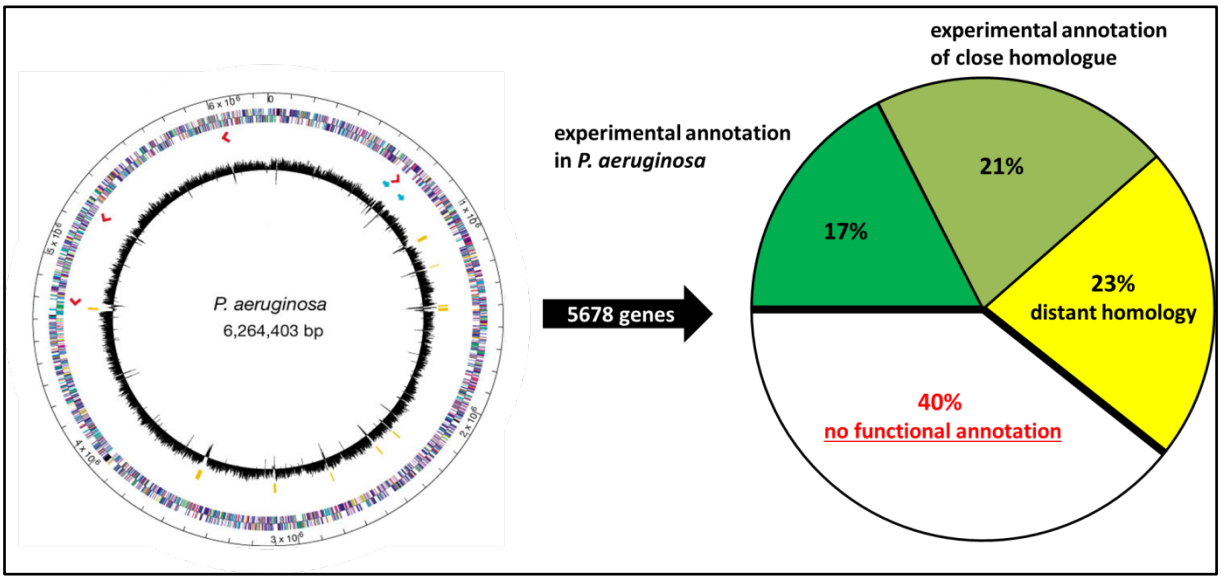


Figure 5: Current status of genomic annotation of *Pseudomonas aeruginosa* PAO1. More than 40% of the 5678 genes (Stover et al., 2000) have not been characterized and are considered as 'white space'. Additionally, another 23% are only annotated according to distant homology (status as of Jan. 21st, 2015).

The problem of lacking functional annotation is recently moving more into the focus with high performance algorithms used to generate information based on primary sequences and predicted structures (Stanberry et al., 2014). This is, without any further experimental evidence, often leading to false annotations (Schnoes et al., 2009), owing to the fact that functional space of proteins is vast and difficult to chart. Current methods are methodologically focused to assign function systematically. For example, if one made an

engineer's deduction that "form follows function", one could conclude that determining the structures of uncharacterized biological macromolecules would reveal their "purpose" (Finkelstein et al., 1993). In fact, one often finds starkly differing function in different members of the same structural family. This is because nature is rarely inventing new folds of proteins but rather uses the existing number of folds as building blocks (Levitt & Chothia, 1976; Islam et al., 1995; Heringa & Taylor, 1997; Apic et al., 2001). Consequently, the *Protein Data Bank* (PDB) is full of entries with 'unknown function', mostly deposited by structural genomics consortia that altogether contribute about 40% of human parasites structures including *Pseudomonas aeruginosa* PAO1 (Chandonia & Brenner, 2006) deposited to the *PDB*. Most of these initiatives are normally not pursuing further experimental characterization. Hence, the Enzyme Function Initiative (EFI) (Gerlt et al., 2011), a collaborative work aims to investigate enzyme functions by an integrated sequence-structure-based approach. Therefore they mostly rely, apart from structural biology, on high-throughput bioinformatics and computational substrate modelling, focusing on enzyme families with known function, so that only its substrate needs to be specified, rendering this approach feasible. However, amongst the number of structures deposited to the *PDB* (317 structures, status as of 21st of January, 2015), only a small fraction was characterized in total.

To fill this information gap, mainly *in-silico* docking is employed to identify possible substrates, which narrows the number of functional hypotheses that needs experimental verification. By using such an approach, limitations are generally introduced. The crystal structure itself is a snapshot of a dynamic protein molecule that might not represent the actual conformation which is adopted while binding its ligand. Furthermore, the physiological relevant interaction partners, the organisms' "metabolome", is not fully characterized, which is thought to contain all metabolites. Its composition cannot be generalized because metabolism is highly strain specific. Until complete characterization, docking calculations rely on either incomplete databases or synthetic compound libraries, irrelevant to the host's metabolome. Therefore, negative results could either be interpreted as a non-small molecule binding protein, or simply that its ligand is not represented in the ligand database used. Moreover, protein conformations determined in the crystal structure and used for docking calculated must not necessary reflect the

conformation which is able to bind any ligand. In this respect, a protein structure can help to explain an experimentally observed function but can normally not be used to derive it.

1.4. A novel approach

To overcome issues that have been mentioned above, a multidisciplinary approach whose final goal is to enlighten a significant part of uncharacterized genomic 'white space' is desired.

Similar approaches however, exclusively focusing on *Pseudomonas aeruginosa PAO1* have been established already, showing the relevance of the bacteria in medicine and biotechnology. These, however, always imply certain restrictions. For example, the AEROPATH initiative (www.aeropath.eu) focuses on genes proven to be essential for infectious disease. Their main objective is structural analysis and drug development whereas the determination of function is not the main focus. The already mentioned EFI focuses only on five enzyme super-families, implying that their general function is already known and the challenge is to assess substrates. Initiatives widely use transposon libraries to elucidate protein function; this however, does not explain the gene's role in genomic context.

Instead of characterizing single genes, the new approach aims to annotate uncharacterized operons first. With the gained information, elucidation of the function of each gene should be easier.

An interdisciplinary approach that combines knowledge gather by structural biology methods with information drafted from metabolomics to derive a hypothetical function of the operon which can be confirmed by biochemical experiments and subsequently leads to new genomic annotations. The general workflow involves bioinformatics, microbiology, molecular biology, metabolomics, x-ray crystallography and biochemistry (figure 6).

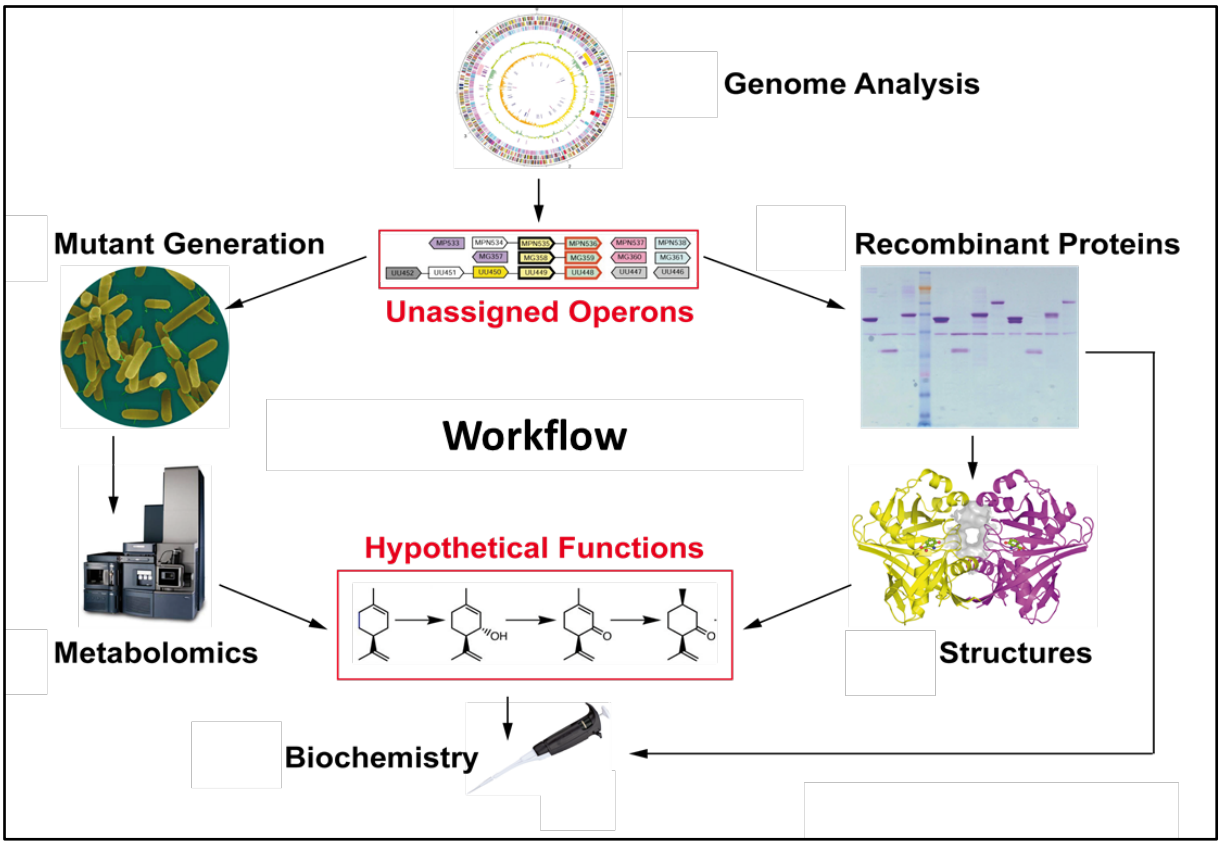


Figure 6: Graphical workflow overview over the approach.

For the left side, covering metabolomics, mutant strains need to be generated in which one of the investigated operons has been deleted or is over-activated. Comparative metabolome analysis of these *lack*-of- and *gain*-of-function mutants might reveal differences and therefore identify a metabolite. This can then be interpreted in the context of the crystal structures of the gen products encoded in the operon. Information from both branches combined will lead to a hypothesis for the function of the operon, which will then be corroborated experimentally. Multiple experimental methods can be employed to identify the function of the single genes. For example, HPLC-MS based analysis can be used to test recombinant proteins for a turnover product, supplying the identified metabolite in purity as substrate. Stepwise protein characterization will lead to a firm final functional annotation of the selected operon.

Once fully established this approach is not exclusively limited to *Pseudomonas* but can, due to its general setup, be applied to investigate other bacterial targets with only minor working effort.

2. Aim of this study

Pseudomonas aeruginosa is one of the most common gram-negative human pathogens in nosocomial infections. The strain PAO1 used in this study comprises a rather large bacterial genome of ~6.5 mega bases, coding for 5678 open reading frames (ORF), of which 40% are not annotated to any specific function. Other than in eukaryotes, bacterial genomes are clustered into operons, with functionally linked genes co-transcribed as one polycistronic mRNA. The genome of *Pseudomonas aeruginosa* PAO1 comprises ~1200 such operons. This simplifies the problem of annotation. By assigning the function of operons first, this information can be used to elucidate the molecular function of the genes included in this operon in a second step. Furthermore, the identification of substrates or products of operons is thought to be easier, as intermediates of physiological pathways are hard to isolate because of their instability and short lifetime.

Therefore, a new structure based approach combined with metabolomic analysis was devised, whose establishment within a pilot-phase is the major aim of this work. For this purpose, two operons were selected and detailed objectives are defined as:

1. Structural characterization of gene products coded in these selected operons
2. Generation of knock-out and gain of function mutant strains as preparation for future metabolome analysis experiments
3. Establishment of a metabolite extraction method
4. Structure-inspired pilot experiments to assess the function of structurally characterized proteins

Combined information from both, detailed structural elucidation and knowledge gain from follow up metabolomic analysis lead to considerable insights into functional space not only in *P. aeruginosa* but also in other microorganisms.

3. Materials

3.1. Chemicals

Chemicals used in this study were purchased in p.a. (*pro analysi*) quality from the following companies: Sigma Aldrich (Hamburg), VWR (Ismaning, Hannover), Merck (Darmstadt), GE Healthcare (Freiburg), New England Biolabs (Frankfurt a. Main) and Qiagen (Hilden).

3.2. Enzymes and Kits

Table 1: Enzymes, marker and kits used in this study

Material	Vendor
Enzymes	
Q5 Polymerase	New England Biolabs
KapaHifi™ Polymerase	PeqLab
Restriction Enzymes	New England Biolabs
TEV-Protease	Lab consumable – self prepared
3C-Precission Protease	Lab consumable – self prepared
T4 DNA Ligase	New England Biolabs
Antarctic Phosphatase	New England Biolabs
Polynucleotide Kinase	New England Biolabs
Taq Polymerase	New England Biolabs
One Taq Polymerase	New England Biolabs
Marker	
SDS-LMW protein standard	Thermo Pierce 26610
DNA ladder	Fermentas Thermo SM0333
Kit	
Plasmid Extraction Kit	Omega
QIAprep Spin Miniprep Kit	Qiagen
QIAquick Gel Extraction Kit	Qiagen
QIAquick PCR Purification Kit	Qiagen
Qiagen Crystal Screening Suites	Qiagen
Midas Screen	Molecular Dimensions
JCSG Core Suite I-IV	Qiagen
Index	Hampton Research

3.3. Bacterial Strains

Table 2: Bacterial strains used in this study; commercial name on the left, corresponding genotype on the right

Strain	Genotype
E. coli XL1blue	*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac*
E. coli BL21(DE3)	*dcm ompT hsdS(rB-mB-) gal*
E.coli BL21(DE3)pLysS	*ompT, hsdSB(rB-rB-), dcm, gal, Cm^r*
E.coli DH5 alpha	fhuA2 lac(del)U169 phoA glnV44 Φ80' lacZ(del)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17
E.coli Rosetta2(DE3) pLysS	F- ompT hsdSB(rB- mB-) gal dcm (DE3) pLysSRARE2 (Cm^r)
Pseudomonas aeruginosa PAO1 wt	
Pseudomonas aeruginosa PAO1 delta PA1621-1624	wt with *del(PA1621-1622-1623-1624)* clean deletion
Pseudomonas aeruginosa PAO1 delta PA3904-3908	wt with *del(PA3904-3905-3906-3907-3908*) clean deletion
Pseudomonas aeruginosa PAO1 delta PA4642-4645	wt with *del(PA4642-4643-4644-4645)* clean deletion

3.4. Media

Table 3: Media composition used in this study; all media were prepared with ddH_2O MilliPore filtered water

Media	Composition
TB – Autoinduction	6g/l Na_2HPO_4, 3g/l KH_2PO_4, 20g/l Trypton, 5g/l yeast extract, 5g/l NaCl, 10ml/l 60% glycerol, 5 ml/l 10% glucose, 25ml/l 20% lactose, 2mM $MgSO_4$
M63	2 g/l $(NH_4)_2SO_4$, 13.6 g/l K_2PO_4, 0.5 mg/l $FeSO_4*7H_2O$, 10 ml/l 20% glucose, 1 ml/l 1M $MgSO_4$
Modified Autoinduction-Super Broth (AI-STB) (Studier, 2005)	35 g/l trypton, 20g/l Yeast extract, 3.3 g/l $(NH_4)_2SO_4$, 6.8 g/l KH_2PO_4, 7.1 g/l Na_2HPO_4, 0.5 g/l glucose 2g/l lactose, 0.15 g/l $MgSO_4$, 5 ml/l glycerol, 0.2x trace elements
Luria-Bertani (LB)	10 g/l tryptone, 10 g/l NaCl, 5 g/l yeast extract
YENB	7.5 g/l yeast extract, 8.0 g/l nutrient broth
SOC	20 g/l trypton, 5 g/l yeast extract, 0.58 g/l NaCl, 0.19 g/l KCl, 10 ml/l 20% glucose, 10 ml 100x $MgSO_4$
Terrific Broth (TB)	12 g/l tryptone, 24 g/l yeast extract, 4 g/l glycerol, 17 mM KH_2PO_4, 72 mM K_2HPO_4
M9	6 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 0.5 g/L NaCl, 1 g/L NH_4Cl, 1 mg/L Thiamine, 1 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.2% (w/v) glucose, 5 g/l glycerol, 0.2 x trace metals
LB Agar	10 g/l tryptone, 10 g/l NaCl, 5 g/l yeast extract, 20 g/l agar

3.5. Antibiotics

Antibiotics were ordered from Sigma and VWR and prepared as 1000x concentrate; ampicillin and gentamycin as 100 mg/ml in H_2O and chloramphenicol as 34 mg/ml dissolved in ethanol. Working concentrations were, unless otherwise specified, 100, 34 and 20 µg/ml for ampicillin, chloramphenicol and gentamycin, respectively.

3.6. Buffer and solutions

Table 4: Buffer composition used in this study

Buffer and solution	Composition
Protein purification	
Buffer A	150 mM Na_2HPO_4, pH 8.0, 300 mM NaCl, pH 8.0
Buffer B	Buffer A plus 500 mM Imidazol, pH 8.0
Buffer C / GF Buffer	50 mM HEPES, 150 mM NaCl, pH 8.0
SDS-PAGE	
4× Stacking Gel Buffer	0.5 M Tris-HCl pH 6.8, 0.4% (w/v) SDS.
4× Separating Gel Buffer	1.5 M Tris-HCl pH 8.8, 0.4% (w/v) SDS.
4× SDS Sample Buffer	130 mM Tris-HCl pH 6.8, 200 mM DTT, 4% (w/v) SDS, 0.025% (w/v) Bromophenol blue, 20% glycerol.
SDS Staining solution	Instant Blue solution
10x SDS running buffer	250 mM Tris, 2 M Glycin, 1% (w/v) SDS.
Agarose gel electrophoresis	
1x TAE buffer	40 mM Tris, 0.1% (v/v) Acetic acid, 1 mM EDTA, pH 8.0
SDS loading buffer	0.03% bromophenol blue, 0.03% xylene cyanol FF, 60% glycerol, 1% (w/v) SDS, 100 mM EDTA (pH 7.6, adjusted by Tris-HCl)
Protease buffer (TEV/3C) A	150 mM Na_2HPO_4, pH 8.0, 300 mM NaCl, pH 8.0
Protease buffer (TEV/3C) B	Protease buffer (TEV/3C) A plus 500 mM Imidazol, pH 8.0
Protease GF buffer (TEV/3C)	50 mM HEPES pH 8.0, 300 mM NaCl, 1 mM TCEP, 10% Glycerol
Trace metal mixture 1000x used at 0.2 x	50 mM $FeCl_3$, 20 mM $CaCl_2$, 10 mM $MnCl_2$, 10 mM $ZnSO_4$, 2 mM $CoCl_2$, 2 mM $CuCl_2$, 2 mM $NiCl_2$, 2 mM Na_2MoO_4, 2 mM Na_2SeO_3, 2 mM H_3BO_3 – dissolved in 60 mM HCl

3.7. Bacterial strains, plasmids and primer

Table 5: General strains, plasmids and primers used in this study

	Description	Reference
Strain		
E.coli DH5alpha	*fhu*A2 Δ(*argF-lacZ*)U169 *pho*A *gln*V44 Φ80 Δ(*lacZ*)M15 *gyr*A96 *rec*A1 *rel*A1 *end*A1 *thi*-1 *hsd*R17	Lab stock (Woodcock et al., 1989)
E.coli XL1	endA1 gyrA96(nal^R) thi-1 recA1 relA1 lac glnV44 F'[::Tn10 proAB$^+$ lacI^q Δ(lacZ)M15] hsdR17(r_{K^-} m_{K^+})	Lab stock
E.coli GC5	FΦ80lacZΔM15 Δ(*lacZYA-argF*)U169 *end*A1 *rec*A1 *rel*A1 *gyr*A96 *hsd*R17 (rk$^-$, mk$^+$) *pho*A *sup*E44 *thi*-1 λ$^-$T1R	Lab stock
E.coli Rosetta 2 (*DE3*) pLysS	F$^-$ *ompT hsdS$_B$* (r_{B^-} m_{B^-}) *gal dcm* (DE3) pLysS*RARE2* (Cam^R)	Lab stock
Pseudomonas aeruginosa PAO1	Wild type PAO1	
Plasmid		
pet19mod	N-terminal His_6-tag, TEV-cleavage site, Amp^R	Lab stock
pLysS	codes for the T7 lysozyme, Cam^R	Lab stock
p5$	N-terminal His_6-tag - TEV cleavage site - T7 lysozyme,	This study
p10$	Expressing N-terminal His_6-tag - TEV cleavage site - T7 lysozyme – 3C cleavage site	This study
pHERDxx	*E.coli/Pseudomonas* shuttle vector	(Qiu et al., 2008)
pEX18Tc	Suicide plasmid for *Pseudomonas aeruginosa*	(Hoang et al., 1998)
Primer 5'-3'		
pLysS_for	5'-TTATAG**CATATG**GCTCGTGTACAGTTTAAAC-3'	This study
pLysS_rev	5'-ATTAA**CATATG**ATCCACGGTCAGAAGTGACC-3'	This study
p5$_delT_for	5'-ACTTCTGACCGTGGA**CATATG**GGACCCGGGGTCGACC-3'	This study
p5$_delT_rev	5'-GGTCGACCCCGGGTCC**CATATG**TCCACGGTCAGAAGT-3'	This study
p5$_point_for	5'-CTTTATTTTCAGGGC**CATACG**GCTCGTGTACAG-3'	This study
p5$_point_rev	5'-CTGTACACGAGC**CGTATG**GCCCTGAAAATAAAG-3'	This study
3C-Prot_for	5'-TATGCTGGAAGTGCTGTTTCAGGGCCCGCA-3'	This study
3C-Prot_rev	5'-TATGCGGGCCCTGAAACAGCACTTCCAGCA-3'	This study
3C_QC_for	5'-GTCACTTCTGACCGTGGA**CATACG**CTGGAAGTGCTGTTTC-3'	This study
3C_QC_rev	5'-GAAACAGCACTTCCAG**CGTATG**TCCACGGTCAGAAGTGAC-3'	This study
pOpine_gfp_ PA1623_f	5'-GAAGTTCTGTTTCAGGGTCCCATGATCGATCTCTACACCGCCG-3'	This study
pOpine_gfp_ PA1623_r	5'-TAAACTGGTCTAGAAAGCTTTAACGGGTCAGCATCGACTGG-3'	This study
pOpine_sumo_ PA5507_f	5'-GCGAACAGATCGGTGGTATGTTCAGCCTGCCCCACC-3'	This study
pOpine_sumo_ PA5507_r	5'-TAAACTGGTCTAGAAAGCTTACAGGTTGCTCCGGCCGAG-3'	This study
	An additional primer list is attached in the appendix.	

4. Methods

4.1. PCR and molecular cloning

4.1.1. Preparation of genomic DNA

Pseudomonas aeruginosa PAO1 genomic DNA was extracted based on a standard protocol from Chen and coworker (Chen & Kuo, 1993) which was slightly adapted. Briefly, 2 ml of a mid-log growth phase bacterial culture grown in LB were harvested by centrifugation at 13,000xg for 2 minutes. Supernatant was decanted and the pellet resuspended in 200 µl of lysis buffer (40 mM Tris-acetate pH 7.8, 20 mM sodium-acetate, 1 mM EDTA, 1% SDS) prior to the addition of 65 µl of 5M NaCl. Cell debris and denatured proteins were removed by centrifugation at 13,000xg for 15 minutes at 4°C. Supernatant was transferred into a new tube and the DNA precipitated by the addition of 400µl of 100% ice cold ethanol and the mixture incubated at -20°C for one hour. Flocculated DNA was separated by centrifugation at 4°C for 20 minutes and was then washed twice by the gentle addition of 200 µl 70% ethanol. Supernatant was carefully removed and pellet dried using speed vac. DNA was resuspended in PCR grade nuclease free ddH_2O to a final concentration of 10 ng/µl, determined by NanoDrop (Thermo).

4.1.2. Extraction of plasmid DNA

5 ml *E.coli* cell culture (*XL1 blue* or *GC5*) harboring the target plasmid was grown in LB media supplemented with appropriate antibiotics overnight. The plasmid extraction from harvested cells was carried out with an alkaline/SDS based lysis using commercially available plasmid Mini-prep kits from either Qiagen or Omega following the vendor's protocol. Bacteria were resuspended in buffer containing agents chelating bivalent cations associated with the cell membrane and therefor destabilizing it. Followed by an alkaline/SDS lysis, the basic pH was neutralized and precipitated genomic DNA as well as lipids and proteins removed by centrifugation at 17000xg. Plasmid DNA remaining in the supernatant was purified by immobilization on a silicon matrix in high salt conditions, washed with ethanol and eluted in low salt with 35-50µl ddH_20.

4.1.3. Polymerase chain reaction (PCR) - *Pseudomonas aeruginosa* genes

In the early days of molecular biology cloning involved the growth of numerous cells from which DNA could be extracted. This became redundant by the invention of the polymerase chain reaction (PCR) in 1986 by Mullis (Mullis et al., 1992), allowing in vitro amplification by using two specific DNA primers, desoxyribonucleotides and a DNA polymerase. Following a three-step temperature protocol, starting with denaturing DNA followed by temperature lowering, generating a DNA hybrid between annealed primer and single stranded DNA. The 3'-OH-primer extension, carried out by the polymerase, yields an exponentially amplified gene defined by the flanking DNA oligonucleotides, during the subsequently carried out reaction cycles.

Generally, a 25 µl PCR reaction contained 5 ng of genomic or 10 ng of plasmid DNA, 20 pmol of specific forward and reverse primer (Table 5, Appendix), 100 µM dNTP mix, 5% DMSO (genomic DNA exclusively) and 2 units of high fidelity polymerase supplied with the corresponding buffer was used. The reaction was carried out in a PeqLab Primus 25 PCR cycler. Initial denaturation was carried out at 95-98°C followed by a 25 second step where primer annealing took place at 56-65°C (see primer table 5, Appendix). The primer extension at 72°C was carried out for 20-180 seconds according to the size of the desired gene. Thirty consecutive repeated steps finalized by a 5 minutes final extension step at 72°C and short term storage at 10°C, yielded sufficient PCR product, which was analyzed with agarose gel electrophoresis using 1% agarose in TAE buffer following standard protocols (Mullis et al., 1992).

Target genes from *Pseudomonas aeruginosa PAO1*: *PA1621*, *PA1622*, *PA1623, PA1624, PA5506, PA5507, PA5508, PA5509* were amplified using appropriate primer sets, designed according to the published genomic sequence of *Pseudomonas aeruginosa PAO1* (Winsor et al., 2011), cross-checked by IDT SciTools (IDT, 2012) and ordered at Eurofins MWG Biotech with HPSF purity (table 5; appendix). A typical PCR protocol is shown in table 6.

Table 6: Typical PCR protocol to amplify from PAO1 genes from genomic DNA

Protocol	Temperature	Time
Prepare PCR mix (25 µl): 10 ng template genomic DNA 3 µl forward primer (10pmol) 3 µl reverse primer (10pmol) 1 µl equal dNTP mix (10mM) 5 µl reaction buffer (5x) 5 µl GC enhancer (5x) Add PCR-grade water to 25 µl	on ice	
PCR-Protocol		
Initial denaturing	95°C	2 min
Cycle denaturing	98°C	20 sec
Annealing	56-65°C	20 sec
Extention	72°C	40 sec/kb
Final extention	72°C	1 min/kb

4.1.4. DNA restriction and cloning

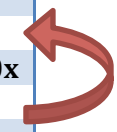
PCR products were purified using QIAquick Gel Extraction Kit or QIAquick PCR Purification Kit (Qiagen) following the supplier's manual. Amplified gene fragments were digested with appropriate restriction enzymes following standard protocols. Briefly, 1 µg of either purified PCR product or plasmid DNA was mixed with 1 µl of both 3' and 5' restriction enzymes supplied in 1x reaction buffer. Digestion of PCR product was carried out at 37°C for 60 minutes, while reactions containing only plasmid were supplied with 1 µl of phosphatase to remove the free 5'-phosphate group to minimize re-ligation of linearized plasmid. Reactions were stopped by the addition of DNA loading dye prior to separation of fragments by agarose electrophoresis. Genetic material was extracted from the gel (QIAquick Gel Extraction Kit), purified and stored at -20°C until used in the ligation reaction. Ligation was carried out in 1x reaction buffer at 16°C for 4 hours. 100 ng of linearized pure plasmid were mixed with a threefold molar excess of desired gene fragment and 500 units of T4-DNA ligase were added. Prior to transformation the ligase was heat-inactivated at 65°C for 25 minutes, as this is a crucial step in efficient bacterial transformation (Ymer, 1991).

4.1.5. Colony PCR and Sequencing

Colony PCR was carried out to screen clones on an agar plate for positive ligation of DNA fragments following standard protocols (Hofmann & Brian, 1991). Briefly, the colony picked from the plate was resuspended in 5 µl H_2O media from which 1 µl was taken as DNA template for PCR and 4 µl were used to inoculate 5 ml LB broth supplemented with appropriate antibiotics. The culture was shaken at 37°C, 180 rpm overnight. PCR reaction was carried out using a plasmid specific flanking primer set, T7 promoter and terminator (table 7). Agarose gel electrophoresis was used to analyze the PCR product. Next day, plasmids from positive clones were isolated and send to either in-house sequencing, using the chain termination method with fluorescently labeled didesoxyribonucleotides (Sanger et al., 1977), or to commercial sequencing (Eurofins MWG Biotech) using both the chain termination and the 454 pyro sequencing method (Binladen et al., 2007; Droege & Hill, 2008).

Table 7: Typical colony PCR protocol using primer set T7 promoter and terminator

Colony PCR-Protocol	Temperature	Time	
Initial denaturing	95°C	2 min	
Cycle denaturing	95°C	20 sec	
Annealing	48°C	25 sec	**30x**
Extention	72°C	45 sec/kb	
Final extention	72°C	1 min/kb	
Storage	4°C	hold	



Plasmids involved in *Pseudomonas aeruginosa* mutant generation were sequenced following a standard BigBye based protocols (Heiner et al., 1998). 100 ng of template DNA were mixed with 4 µl BigDye (Life Technologies) and 1 µl forward or reverse sequencing primer. The reaction was carried out with initial denaturation at 94°C for 2 minutes. Following 60 cycles of denaturation for 30 seconds at 94°, primer annealing at 55°C (20 seconds) and primer extension for 4 minutes at 65°C, the reaction products were stored at 10°C. Excessive nucleotides solution as well as primer and polymerase were removed by DTR gel filtration cartridges (EdgeBio, Gaithersburg, MD 20877, USA) following the vendor's protocol.

4.1.6. QuikChange mutagenesis

The QuikChange™ reaction following standard protocols (Braman et al., 1996; Liu & Naismith, 2008) was used for side-directed mutagenesis to introduce point mutations. Two primers a 125 ng, each complementary to the opposite DNA strand, were mixed with a template plasmid.

$$X\ pmoles\ of\ oligo = \frac{125\ ng\ of\ oligo}{330 * number\ of\ bases\ in\ oligo} * 1000$$

Equation: Calculation of primer amount

The primers were extended during the PCR cycle, resulting in a DNA double strand. The hybrid strand consisted of one parental wild type strand and one daughter strand, carrying the desired mutation. The number of consecutively carried out cycles was varied between 12 and 18 according to the mutation desired (Table 8). Recognizing and digesting only methylated DNA at the target sequence 5-GA_mTC, 20 units of *DpnI* were added to digest parental DNA. Appropriate negative controls were transformed side by side with the QuikChange™ reaction mixture into competent cells, incubated and plated on LB-agar containing selecting antibiotics.

Table 8: Cycling parameters according to the type of mutation desired

Type of mutation desired	Number of cycles
Point mutations	12
Single amino acid change	16
Multiple amino acid deletions or insertions	18

4.2. Bacterial transformation - general aspects

Bacteria can actively take up external plasmid DNA and either harbor the plasmid or heritably incorporate it into the genome (Griffith, 1928; Downie, 1972). Genetic diversity and exchange of genetic information is a dynamic process involving uptake and active release of DNA containing vesicles (blebs) within a culture of gram-negative bacterial species including *Escherichia coli* and *Pseudomonas aeruginosa* (Dorward & Garon, 1990), both used in this study. The presence of calcium was shown to be important to bacterial competence and essential for uptake and exchange of genetic material (Cohen et al.,

1972, 1973; Page & Doran, 1981; Trombe et al., 1992; Trombe, 1993). Nowadays two methods of bacterial transformation have been established: electro transformation, which requires high cell densities and special equipment, and the chemical transformation, a procedure which requires only a temperature controlled water bath.

4.2.1. Electro competent cells

Electro competent cells were used in this study for plasmid transformation to either expression strains or cloning cell lines. Preparation was modified as reported elsewhere (Gonzales et al., 2013). Briefly, a fresh cell culture in 40 ml YENB Medium was grown in a 500 ml flask overnight. The next day, 1 liter YENB media in a 2 L flask was inoculated with 10 ml of the overnight culture, incubated at 37°C and shaken. Cells were harvested in sterile chilled centrifuge bottles at an OD_{600} of 0.45 (Tu et al., 2005) and immediately placed on ice. Supernatant was decanted and pellet resuspended in 10 ml sterile fresh autoclaved MilliQ water-cooled to ice temperature to wash the cells twice. The supernatant of the final wash was removed and the pellet washed with 20ml ice-cold sterile 15% glycerol once, pelleted and carefully resuspended in 10 ml ice-cold sterile 15% glycerol. Aliquoted into 50 µl portions, they were either used right away or flash frozen and kept at -80°C for long-term storage.

4.2.2. Chemical competent cells

Chemical competence of *Escherichia coli* strains was achieved by a $CaCl_2$ based protocol (Inoue et al., 1990). Briefly, a one liter LB culture was inoculated 1:100 from 50 ml stationary phase overnight culture and grown at 37°C, 200 rpm until an OD_{600} of 0.3 (Tu et al., 2005). Cells were harvested and first resuspended in approximately ½ pellet volume of ice-cold 100 mM $CaCl_2$, incubated for 10 minutes, harvested and resuspended again in 1/10 volume of ice-cold 100 mM $CaCl_2$. After one hour incubation on ice, 50 µl aliquots were either immediately used or glycerol was added to a final concentration of 15% prior to flash freezing and long term storage at -80°C.

4.2.3. Transformation of genetic material

Electroporation was performed in a 1 mm cuvette applying an electric pulse at 25 µF, 200 Ω and 1.8 kV for 5 ms (*E.coli* Pulser, Biorad), generating an electric field of 18000 V/cm^{-1} as reported elsewhere (Calvin & Hanawalt, 1988). About 5 ng of purified plasmid DNA were mixed with 50 µl of electro competent cells on ice and transferred to a chilled cuvette. After electroporation, 1 ml of pre-warmed LB or SOC media was added and the cells incubated for 1 hour at 37°C. Transformed clones were selected on LB-agar supplied with the appropriate antibiotic. Transformation of the product of a ligation reaction required an additional desalting step via micro dialysis as elsewhere described (Saraswat et al., 2013).

Chemical transformation was accomplished by heat shock treatment (Froger & Hall, 2007). 10 ng of plasmid DNA was added to an aliquot of competent cells and incubated for 30 minutes on ice. Heat shock was achieved by transferring the tubes to 42°C water bath for 45 seconds. Pre-warmed, 1ml SOC media was added and the cells were incubated for 1 hour. Transformed clones were selected the same way as described above.

4.2.4. Bacterial storage

Both *Escherichia coli* and *Pseudomonas aeruginosa PAO1* were grown in LB media with appropriate antibiotics if needed. Once they reached an OD_{600} of 0.5, equal volumes of bacterial broth and 80% sterile glycerol were mixed in a cryogenic tube and flash frozen in liquid nitrogen prior to storage at -80°C.

4.3. Expression plasmids – general aspects

Forced high-level heterologous protein expression in bacteria can result in large amounts of incorrectly folded protein (Hammarström et al., 2002). Aggregated protein may be easy to purify, however, obtaining active protein then typically requires protein-specific and labor-intensive *in vitro* re-folding steps with no guarantee of biologically active product (Singh & Panda, 2005). To avoid protein aggregation, efficient and well established solubility tags (Young et al., 2012) like maltose binding protein (MBP) (di Guan et al.,

1988) and glutathione-S-transferase (GST) (Smith & Johnson, 1988) as well as smaller tags like SUMO (Panavas et al., 2009) and a new generation Strep-tag (Schmidt et al., 2013) were developed. Thus, with the increasing demand for novel expression techniques, expression plasmids featuring novel solubility helper are desirable. In this work, tailor made expression plasmids based on the *pET*-system (Studier & Moffatt, 1986) (Novagene), suitable for recombinant protein expression in lysogenic *Escherichia coli* strains, were developed and successfully used.

4.3.1. Generation of *p5$*

The gene coding for T7-lysozyme was PCR amplified from the *pLysS* plasmid using the primer set 5'-TTATAGCATATGGCTCGTGTACAGTTTAAAC-3' and 5'-ATTAACATATGATCCACGGTCAGAAGTGACC-3'. The resulting 0.46 kbp product was gel purified, digested with *NdeI* and ligated into the single cut side of the plasmid *pet19mod*. Ligation mixture was transformed into chemical competent *XL1 blue* cells and SOC media (1ml) added for cell recovery. After incubation for 50 minutes at 37°C, cells were plated on ampicillin containing agar plates. The next day, all clones were scraped of the plate, combined in 5 ml LB broth and incubated at 37°C, 180 rpm. After four hours, plasmids were recovered (Qiagen Mini Prep-Kit) and 1 µl of the plasmid mixture was used as template in a subsequent Quik-Change (Stratagene) reaction. The primer set used in this reaction not only introduced a mutation at the first 5'-NdeI recognition site in order to only leave the 3' original *NdeI* cloning side intact but it was also designed such that only plasmids carrying correctly inserted T7-lysozyme genes were amplified. *DpnI* digestion following supplier's protocol was carried out to digest parental plasmid DNA prior to transformation into fresh *XL blue* cells. Cells were spread on agar plates and incubated at 37°C overnight. Single clones were picked, grown in 5ml LB media, and the plasmid was extracted after 10 hour incubation at 37°C, 180 rpm. DNA sequencing (Eurofins MWG Biotech) using primers recognizing the *T7_prom* and *T7_term* region (Mead et al., 1986) verified correct insertion of the gene and successful mutagenesis (figure 7).

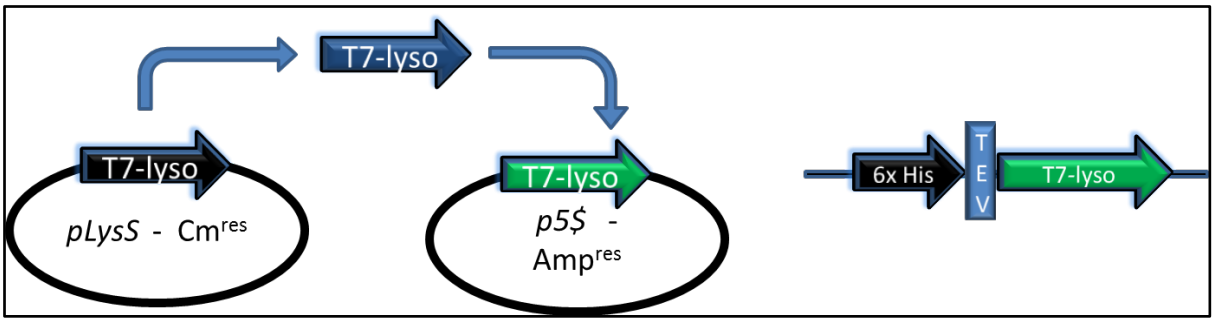


Figure 7: Scheme of the construction of *p5$*.

4.3.2. Engineering of *p10$*

The plasmid *p5$* was further engineered to the more advanced plasmid *p10$*. Two oligo nucleotides coding for the full length Prescission protease recognition site (LeuGluValLeuPheGln|GlyPro; | = indicating the cleavage site) (Cordingley et al., 1990) were ligated into *p5$* linearized with *NdeI.* First, plain primers were phosphorylated by polynucleotide kinase following the vendor's manual. Namely, a 30 µl reaction contained 150 pmol of either primer, polynucleotide kinase buffer, 1 mM ATP and 5 units of enzyme was incubated at 37°C for 30 minutes prior to heat inactivation at 65°C for 30 minutes. Phosphorylated primers were mixed in a 1:1 ratio and heated to 95°C for 10 minutes using a PCR cycler. Slow primer annealing was performed by walking down a linear temperature gradient to 20°C over a period of 60 minutes. The reaction mixture was stored at -20°C until use. The template *p5$* was linearized using the single cut side *NdeI* at which the annealed primer dimer was inserted using regular ligation reaction. After heat inactivation of the ligase the reaction mixture was desalted by micro-dialysis on a nitrocellulose filter floating on ddH_2O for 20 minutes. Half of the recovered volume was transformed into electro competent *XL1 blue* cells and spread on ampicillin-containing agar and incubated at 37°C overnight. Next day, cells were harvested from the LB plate, resuspended in 4 ml LB media, incubated for six hours and the plasmids were extracted. Secondly, 1 µl of the plasmid composite was used in a Quik-Change (Stratagene) reaction to remove the 5'-*NdeI* site with an appropriate primer set (table WWW). Digestion of parental plasmid was carried out by *DpnI,* followed by transformation into chemical competent *DH5 alpha* cell and cells were plated on LB agar. Plasmids were recovered from single colonies. DNA sequencing using T7_prom and T7_term primer verified correct insertion of the primer dimer as well as accurate mutagenesis.

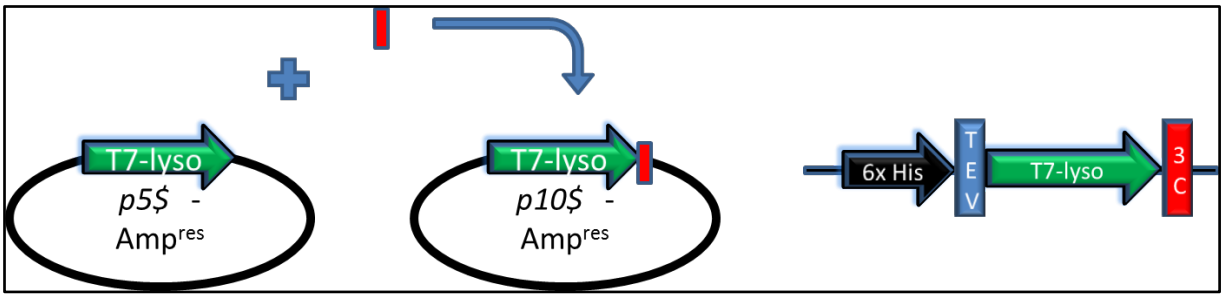


Figure 8: Schematic representation of the construction of *p10$* from *p5$*

4.3.3. Construction of *pCryst* and *pUrgé*

The *T4-lysozyme* gene was PCR amplified from the plasmid *pT4_lysozyme_D20A* (addgene: 18241; www.addgene.org). Two primer sets were used to generate PCR products with either *NdeI* or *XhoI* restriction sides. The PCR product was analyzed by gel electrophoresis and the band at 0.60 kbp corresponding to the amplified gene was cut, gel purified, and digested with appropriate restriction enzymes *NdeI* or *XhoI*, respectively. The same procedure as described for construction of *p5$* was followed. PCR product insertion into the linearized plasmid *pet19mod* using the *NdeI* recognition site followed by a subsequent QuikChange reaction, resulted in the plasmid *pCryst*. Repeated insertion at the *XhoI* side of linearized *pCryst* also followed by a QuikChange reaction yielded the sandwich-plasmid *pUrgé*. Correct integration and mutagenesis was verified by DNA sequencing (Eurofins MWG Biotech).

4.3.4. Plasmids constructed and used in this work

Table 9: List of plasmids: Inserted genes originated from *Pseudomonas aeruginosa PAO1*

Plasmid	Inserted gene	Antibiotic resistance
p10$_PA1621	*PA1621*	Ampicillin
p10$_PA1622	*PA1622*	Ampicillin
pOPINE_GFP_PA1623	*PA1623*	Ampicillin
p10$_PA1624	*PA1624*	Ampicillin
pET19mod_5506	*PA5506*	Ampicillin
pOPINE_SUMO_PA5507	*PA5507*	Ampicillin
p10$_PA5508	*PA5508*	Ampicillin
p10$_PA5509	*PA5509*	Ampicillin

4.4. Gene expression and protein purification

4.4.1. General aspects and plasmid design

Heterologous recombinant protein expression and purification is the heart of protein crystallography and biochemical studies of proteins. Versatile organisms have been implemented but the number of significantly used hosts can be narrowed down to a few (Figure 9). The system of choice in this study was the well characterized and easy to genetically manipulate *Escherichia coli*.

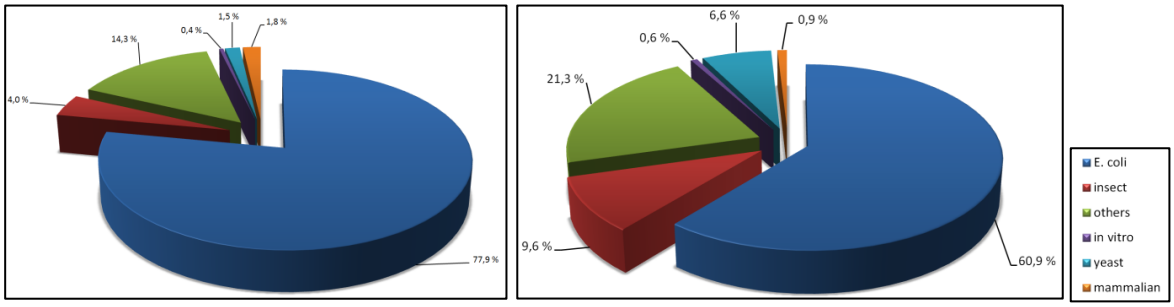


Figure 9: Usage of host cell lines for expression of cytosolic (left) and integral membrane proteins (right) for following structural studies 2004-2013; source: protein data bank (Berman et al., 2000).

Over the last decades, ingenious *E.coli* genetic engineering was arising side-by-side with developments in cloning and plasmid design. The *pMB1* origin, present in all *pET* plasmids (Novagene), restricts the plasmid copy number to 15-60 per cell (Bolivar et al., 1977), while slight mutations in *pMB1* resulting in the *pUC* origin (*pHERD* plasmids) increased the number up to 700 copies / cell (Minton, 1984). The combination of replicon and promoter strength dictates the number of translatable polynucleotide templates. The promoter used in this study was the *T7 promoter* (Mead et al., 1986) derived from the *T7 phage* in combination with the phage RNA polymerase (T7RNAP). The gene coding for the DNA dependent RNA polymerase is commonly integrated in the *E.coli* genome under the control of a *lac*UV5 promoter, ten times stronger than wild type *lac* promoter (Amouyal & Buc, 1987).

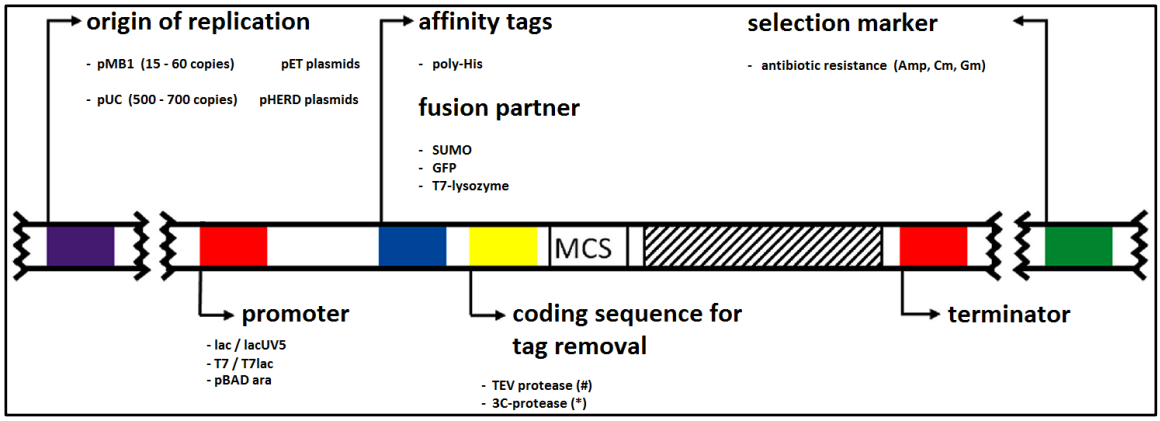


Figure 10: Schematic architecture of an expression plasmid, depicting major features in common expression vectors.

To avoid leaky expression the strains BL21*(DE3)pLysS* or Rosetta2*(DE3)pLysS* were used in this study. These cells carry the plasmid *pLysS*, from which a natural inhibitor of T7RNA polymerase, the T7 lysozyme, is constitutively expressed, which binds to T7RNAP and inhibits transcription initiation (Studier, 1991; Jeruzalmi & Steitz, 1998; Cheetham et al., 1999). The expression of the target gene from the *pET*-based plasmids was either induced by the addition of lactose (Studier, 2005) or a non-hydrolysable analogue, isopropyl β-D-1-thiogalactopyranoside (IPTG) (Hansen et al., 1998). This yielded soluble and pure protein, which was the major requirement for this study. The expression temperature (Hewitt & McDonnell, 2004) as well as media selection were shown to have a major impact on the final protein yield (Studier, 2005; Sivashanmugam et al., 2009). To identify the best combination of both, all constructs used in this study have been test-expressed using different media and various expression temperatures. Besides refining these physical external factors, genetic engineering was used to overcome issues of imbalanced codon usage between gene donor organism (*Pseudomonas aeruginosa*) and expression host (*Escherichia coli*), which could result in premature termination of translation. For instance the second triplet following the start codon, shown to have up to 15-fold influence on gene expression level (Looman et al., 1987), was optimized for all plasmids used in this study. To enhance protein expression level and solubility, the SUMO protein (Butt et al., 2005; Panavas et al., 2009) as well as a variant of the green-fluorescent-protein (GFP) (Murayama & Kobayashi, 2014) and the T7-lysozyme fusion, developed in this study, were employed as genetically fusion partners. All fusion protein featured the His_6-affinity tag for affinity chromatography.

4.4.2. Native protein expression

In all cases, target genes were expressed in various *E. coli* strains. A single colony of plasmid- harboring cells was used to inoculate a 50 ml overnight culture supplemented with appropriate antibiotic(s) and continuously shaking at 200 rpm at 37°C. The next day, fresh medium was inoculated at an OD_{600} of 0.02. Cells were grown in media supplemented with antibiotics to an OD_{600} of 0.7 at 37°C, then the temperature was lowered to 20°C and 0.5 mM IPTG was added to induce protein expression. Cultures were shaken for another 14 hours. If auto-induction was performed, plasmid harboring cells were grown in a modified auto induction medium (AI-STB) supplemented (Studier, 2005) with antibiotics to an OD_{600} of 0.5 at 37°C before the temperature was lowered to 25°C and the cultures were shaken for another 36 hours. Cells were harvested by centrifugation for 15 minutes at 6500xg. If not immediately processed further, the cell pellet was transferred into a falcon tube, flash frozen in liquid nitrogen and stored at -80°C until further use. The following table 10 summarizes the expression conditions for the different genes.

Table 10: Bacterial expression strains and plasmids used in different expression conditions

Protein	E.coli strain	Plasmid	Medium	IPTG	Temperature	Time
PA1621	Rosetta2*(DE3)pLysS*	*P10$*	LB	0.5 mM	37°C -> 20°C	16 hours
PA1622	Rosetta2*(DE3)pLysS*	*p10$*	TB	0.5 mM	37°C -> 20°C	14 hours
PA1623	BL21*(DE3)pLysS*	*pOPINE_GFP*	TB	0.5 mM	37°C -> 20°C	12 hours
PA1624	BL21*(DE3)pLysS*	*p10$*	LB	0.5 mM	37°C -> 20°C	14 hours
			AI-TB		37°C -> 25°C	36 hours
PA5506	BL21*(DE3)pLysS*	*pET19mod*	TB	0.5 mM	37°C -> 20°C	14 hours
PA5507	Rosetta2*(DE3)pLysS*	*pOPINE_SUMO*	AI-STB	none	37°C -> 25°C	36 hours
PA5508	BL21*(DE3)pLysS*	*p10$*	LB	0.5 mM	37°C -> 20°C	14 hours
PA5509	Rosetta2*(DE3)pLysS*	*p10$*	LB	0.5 mM	37°C -> 20°C	12 hours

4.4.3. Expression of selenomethionine labeled protein

Both, L-Seleno-methionine and L-seleno-cysteine is suitable replacements for their natural amino acids to perform anomalous diffraction experiments in protein crystallography. In this study L-seleno-methionine was incorporated by using minimal media (M9) as described by Studier (Studier, 2014). Test expressions revealed Rossetta2*(DE3)pLysS* and BL21(*DE3)pLysS* as the best hosts for labeled protein expression.

A single colony harboring the expression plasmid was used to inoculate a 100 ml LB media overnight culture supplemented with antibiotics, incubated at 37°C, 180 rpm. The next day, cells were harvested and resuspended in 150 ml M9 media with proper antibiotics and incubated for 1 hour further, while fresh media for the expression culture was pre-warmed to 37°C and pre-aerated at 130 rpm. Main cultures were inoculated with an OD_{600} of 0.01 and continuously shaken at 130 rpm, 37°C. The optical density was monitored every hour until it reached 0.5. An amino acid mix to inhibit natural methionine biosynthesis was added (100 mg/l K, F and T; 50mg/l I, L and V) and growth continued for an additional 20 minutes, while temperature was lowered to 20°C or 25°C after 10 minutes. Protein expression was induced by the addition of 0.5 mM IPTG and simultaneously 60 mg/l L-selenomethionine was added to be preferentially incorporated into the overexpressed protein. Constant shaking at 20°C or 25°C was carried out for 12-18 hours. Details of labeled protein expression are shown in the table 11.

Table 11: Expression details of selenomethionine labeled protein

E.coli strain	Protein	Plasmid	Media	IPTG	Temperature	Time
Rosetta2*(DE3)pLysS*	PA1622	*p10$*	M9	0.5 mM	37°C -> 20°C	16 hours
BL21*(DE3)pLysS*	PA1623	*pOPINE_GFP*	M9	0.5 mM	37°C -> 20°C	12 hours
BL21*(DE3)pLysS*	PA1624	*p10$*	M9	0.5 mM	37°C -> 25°C	12 hours
BL21*(DE3)pLysS*	PA5506	*pET19mod*	M9	0.5 mM	37°C -> 20°C	18 hours

4.5. Protein purification

4.5.1. Metal affinity chromatography – general aspect

Transition metals show a natural affinity to histidine residues embedded in a polypeptide chain. Based on this, the concept of immobilized metal affinity chromatography (IMAC) was developed (Porath et al., 1975). In this study IMAC was used as a first step protein purification from cleared lysate using the His_6-tag (Hochuli et al., 1987) of the recombinantly expressed proteins (Figure 11).

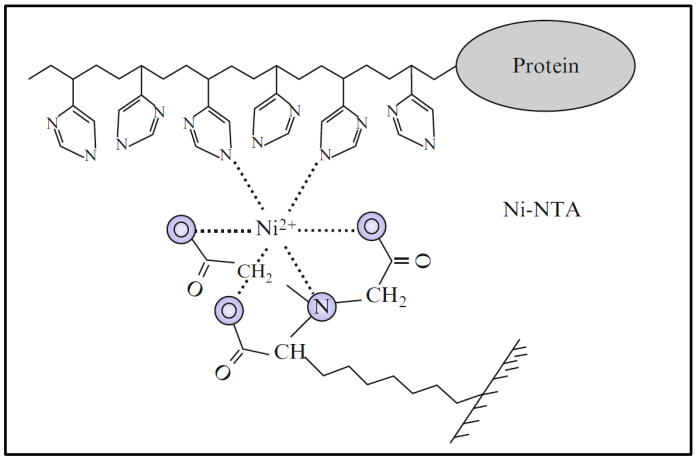


Figure 11: Interaction model of a His_6-tagged protein with a NTA modified resin. Nickel is coordinated in tetradentate coordination number four by the resin, leaving two valences for interaction with the imidazole ring of the histidine side chains of the His_6-tag.

4.5.2. Nickel affinity chromatography

Cells were harvested by centrifugation and then resuspended in buffer A containing 100 µM of the protease inhibitor phenylmethylsulfonyl fluorid (PMSF). Lysis was performed by passing the cells 5 times through a Microfluidizer (Microfluidics) or Avestin Homogenizer. Cell debris and insoluble matter were separated from the supernatant by ultracentrifugation at 35,000 g for 1 hour at 4°C before the cleared lysate was loaded on a HiTrap Chelating HP column (GE Healthcare) pre-charged with 100 mM $NiSO_4$ and equilibrated with buffer A, using an ÄktaPrime System (GE Healthcare). The column was then washed with buffer A until the OD_{280} remained constant. Non-specifically bound proteins were removed with 2% buffer B (10 mM imidazole) until the OD_{280} was constant. A gradient to 100% buffer B over a volume of 240 ml was used to elute specifically bound target proteins. Fractions containing target protein were identified by standard SDS-PAGE gel electrophoresis and pooled together. The protein concentration of pooled fractions

was measured by UV-absorption using NanoDrop (Thermo) and the protein-specific extinction coefficient. Protease was added to remove affinity tags during dialysis (Snake Skin, Thermo, 10 kDa cut off membrane) against buffer GF at 4°C or 20°C overnight, depending on the protein purified. Precipitated protein was removed by centrifugation (5000 g, 20 minutes) and the supernatant was passed over a HiTrap Chelating HP column (GE Healthcare) pre-equilibrated with buffer GF to separate uncleaved fusion protein from tag-free target protein. A gradient from 0% to 100% buffer D was performed to elute tagged-protease as well as uncleaved protein from the column. The flowthrough and the collected portions were analyzed by SDS-PAGE, fractions containing target protein were pooled and concentrated by ultracentrifugation (Pall Corporation, Macrosep® advance, 10 kDa) for further purification steps. A summary of purification details is given in table 12.

Table 12: Details for all native expressed and purified proteins in this work

Protein	Molecular weight uncleaved / cleaved [Dalton]	Protease	Dialysis temp. [°C]	Extinction coefficient uncleaved / cleaved [$mol^{-1}cm^{-1}$]
PA1621	50733 / 30471	Precission	---	55140 / 26930
PA1622	51552 / 31290	Precission	20	55140 / 26930
PA1623	53760 / 24924	Precission	4	57425 / 35410
PA1624	48015 / 27753	Precission	4	59150 / 30940
PA5506	50733 / 30471	TEV	20	55140 / 26930
PA5507	36617 / 24262	SUMO	4	20190 / 18700
PA5508	68162 / 47900	Precission	4	78225 / 50015
PA5509	45103 / 24841	Precission	---	58120 / 30035

4.5.3. Size exclusion chromatography

Size exclusion chromatography (SEC), also called gel filtration (GF), is a technique in liquid interaction chromatography (LIC). Porous material separates particles due to their diverse interaction and slightly different diffusion path through the matrix material. This results in varied retention volume (V), reflected by the interstitial (V_i) and the pore volume (V_p) of the matrix phase as well as a distribution coefficient K (Gorbunov & Skvortsov, 1995).

$$V = V_i + K * V_p \quad with \quad K = 1 + {}^{2R}/_{D}\left[\frac{Y(-cR)-1}{cR} - {}^{2}/_{\sqrt{\pi}}\right]$$

The distribution coefficient K is defined by the analyte radius of gyration R, the average pore diameter of the matrix material D and a general adsorption parameter c (Gorbunov & Trathnigg, 2002). Generally, if K < 1 the analyte molecules cannot be resolved and elutes before the void volume, this is typical the case for aggregates. For K > 1 the analyte molecules can be separated on the column resulting in an elution profile (figure 12). SEC was used in this work to carry out the final purification step as well as to determine the oligomeric state of a purified protein.

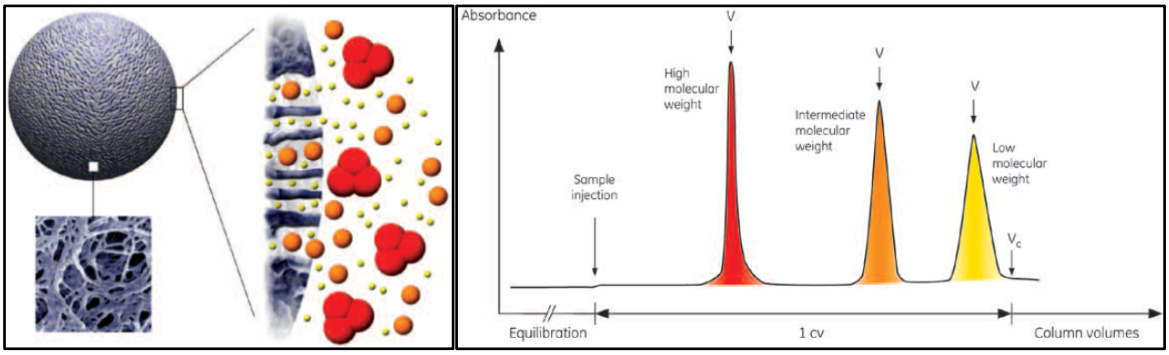


Figure 12: Porous material of SEC columns and schematic interaction of analyte mixture with the SEC material – resulting in a different column retention volume, right (adapted from (GE, 2010))

A concentrated protein solution of up to 12 ml was injected onto a size exclusion chromatography column HiLoad Superdex S75 (16)26/60 or HiLoad Superdex S200 (16)26/60 (GE Healthcare) pre-equilibrated to buffer D using an ÄktaPurifier System (GE Healthcare). An isocratic elution at a flow rate of 1-2 ml/min was performed and the UV trace at 280 nm was followed. Collected fractions of 4 ml were analyzed upon its protein content by SDS-PAGE. Fractions containing pure target protein were pooled and further concentrated via ultrafiltration (Pall Corporation, Macrosep® advance, 10 kDa) for crystallization experiments. The protein‘s molecular weight was calculated from its retention time using the middle of the UV peak against a standard calibration curve, generated by a protein mixture of known size. Size exclusion chromatography was performed at either 4°C or 10°C, as depicted in detail in table 13.

Table 13: Size exclusion chromatography details

Protein	Molecular weight cleaved [Dalton]	Gelfiltration temperature [°C]	Extinction coefficient cleaved [$mol^{-1}cm^{-1}$]	Gelfiltration column (GE Healthcare) Superdex HiLoad
PA1621	30471	---	26930	---
PA1622	31290	10	26930	S26/60 75
PA1623	24924	4	35410	S26/60 75
PA1624	27753	10	30940	S26/60 75
PA5506	30471	10	26930	S26/60 200
PA5507	24262	4	18700	S26/60 75
PA5508	47900	10	50015	S26/60 200
PA5509	24841	---	30035	---

4.5.4. TEV protease

TEV protease is the name for the 27kDa catalytic domain of the nuclear inclusion a protein (Nia) encoded by the tobacco etch virus (TEV). Because of its stringent sequence specificity it is commonly used as a reagent for cleaving recombinantly expressed fusion protein. TEV protease recognizes a linear epitope of the general form of E- X_{aa} - X_{aa} -Y- X_{aa} -Q-|-G/S with | indicating the proteolytic cleavage site. Some variations at position P5, P4 and P2 are enzymatically tolerated, however, even conservative mutations at position P2 (F to Y) and P4 (L to F) reduce the processing efficiency by two orders of magnitude. The crystal structure solved in complex with an oligopeptide substrate (Phan et al., 2002) revealed that only the amino acid at P5 does not intimately contribute to substrate binding and can therefore be diversified. Systematic study did demonstrate that an amino acid variation at the P1' position results only in a moderate change in catalytic efficiency of the protease. The S1' subsite is not a true pocket but rather a shallow groove with no direct substrate specificity (Kapust et al., 2002). However, the most efficient sequence was ENLYFQ|S.

The drawback of commercial TEV protease (LifeTechnology), apart from its high cost, is that it cleaves itself after time, thus specific proteolytic digestion of target fusion protein is restricted to a certain time frame. Therefore a mutant of TEV-protease (S219V)

reported to be imperviously resistant to autolysis (Kapust et al., 2001) was employed in this study. The gene coding for the protease was genetically fused to maltose binding protein (MBP) increasing expression level by orders of magnitude. Sophisticated plasmid design was carried out by introducing an additional TEV-epitope recognition site. The fusion protein is undergoing post-translational auto-processing by cleaving MBP from itself, liberating the featured His_6-tag which is needed for follow up protein purification. As a result of high expression level, fast folding of TEV-fusion-protein induced by MBP, high yields of pure protease could be obtained.

In detail, 4 liters of LB media supplemented with appropriate antibiotics was inoculated with 10 ml of a pre-culture grown from a single colony of plasmid harboring *E.coli* BL21*(DE3)pLysS*. Cells were shaken at 37°C until OD_{600} of 0.5 was reached, temperature was lowered to 20°C, IPTG added at 0.5 mM and further incubation was carried out for 10 hours. Cells were harvested and resuspended in buffer A. After cell lysis, insoluble matter was separated by ultracentrifugation and soluble fraction was loaded onto a nickel affinity column (GE Healthcare) pre-equilibrated to buffer A. Column was washed thoroughly with buffer A supplemented with 2% buffer B until UV trace reached baseline after washing with 50 ml. A continues gradient to 60% buffer B over 240 ml was performed to elute the specifically bound protease. Fractions containing pure target protein were identified by SDS-PAGE, pooled and dialyzed against protease buffer (50 mM HEPES, 300 mM NaCl, 1mM TCEP, 10% glycerol) overnight. Next day a size exclusion chromatography was performed using a 26/60 S75 column (GE Healthcare) pre-equilibrated to the same buffer. Peak fraction contained pure protein were pooled and concentrated to a maximum of 3 mg/ml using ultra centrifugation (Pall MacroSep, 10kDa). 500µl aliquots were flash frozen in liquid nitrogen and stored at -80°C until use. Each batch was tested and the amount used adjusted, thus working concentrations of 1:30 were used on average for processing target fusion protein during dialysis.

4.5.5. Rhino-Virus 3C Protease

Human rhinovirus (*HRV*), a member of the *picornaviridae*, is the major agent causing the common cold in humans with worldwide health and economic impact (Jensen et al., 2015). Its small positive-stranded RNA genome, about 7200 nucleotides in size (Stanway et al., 1984), encodes two viral proteolytic enzymes, $2A_{(pro)}$- and $3C_{(pro)}$-protease within the single viral 243 kDa poly-protein (Callahan et al., 1985). The activity of both proteases, is crucial for virus replication by processing the large viral precursor poly-protein as well as digestion of cellular proteins, renders them a viable drug target (Baxter et al., 2011). In fact, 3C protease was already successfully targeted in treatment of common cold in experimental phase (Matthews et al., 1999). Structural characterization of the protease (Matthews et al., 1994) revealed a trypsin-like fold and active site composition but possessing characteristics of both, serine and cysteine protease with the catalytic triad Cys-146, His-40 and Glu-71. The protease recognizes an epitope stretch of minimum six amino acids (TLFQ|GP; | indicates the cleavage site) (Cordingley et al., 1990). Studies on amino acid composition revealed that the peptide recognition pattern of EVLFQ|GPVY showed the highest catalytic activity and very conservative mutations on either P4 or P5 dropped K_{cat}/K_m by orders of magnitude (Cordingley et al., 1990). This could be explained by the crystal structure solved (Matthews et al., 1994), however, the common used recognition pattern is LEVLFQ|GP showing highly efficient cleavage. In this study the 3C protease of HRV-14 was employed to cleave recombinant expressed fusion protein at its incorporated specific recognition site.

Briefly, the gene coding for the HRV-14 3C-protease was expressed in plasmid harboring BL21*(DE3)* cells grown in TB media at 37°C. Temperature was lowered to 20°C and protein expression induced by the addition of 0.5 mM IPTG at OD_{600} of 0.8 and further shaken for 12 hours. The His_6-tag featuring recombinant protein was purified from cleared bacterial lysate in a 2-step manner including both standard metal affinity- and size exclusion chromatography as described in detail for TEV-protease before. Pure protein was concentrated to 2.5 mg/ml, aliquoted into 500µl fractions and immediately flash-frozen in liquid nitrogen prior to storage at -80°C.

4.5.6. SUMO protease

SUMOylation of recombinant expressed protein is commonly used to overcome major issues of recombinant protein expression and insolubility (Panavas et al., 2009). A SUMO-specific protease removes the small-ubiquitin modifier (SUMO) from a target protein and facilitates the generation of protein of interest with a native N-terminus from an expressed SUMO-fusion protein. Other than the previously described TEV- or HRV-3C-protease SUMO protease does cleave at a specific epitope sequence but rather recognizes the typical β-grasp fold of the SUMO protein (Hochstrasser, 2009). In this study the highly active cysteinyl protease also known as Ulp, a recombinant fragment of the ubiquitin-like-specific protease 1 (Ulp1) from *Saccheromyces cerecisiae*, was used (Li & Hochstrasser, 1999). This enzyme cleaves the peptide backbone between the terminal double glycine pair and the adjacent amino acid of the fusion protein which is mimicking its natural C-terminal extended precursor form (Hickey et al., 2012). The optimal temperature for cleavage is 30°C; however, the enzyme is active over a wide range of temperature and pH (pH 7.0-9.0). Preparation was carried out according to standard procedure. Briefly, the SUMO protease was expressed as a His_6-Tag fusion protein in *E.coli* BL21*(DE3)pLysS* grown in LB broth and purified in a two-step manner from cleared lysate as described previously. The pure 26 kDa protein was aliquoted, flash frozen and stored at -80°C. It was used in a 1:100 ratio to cleave SUMO-fusion protein during dialysis.

4.5.7. Purification of PncA - Nicotinamidase

The protein was purified as described previously (Garrity et al., 2007). In short, pPNC2 harboring BL21(*DE3)pLysS* cells were grown in lysogenic broth supplemented with appropriate antibiotics. Cultures shaking at 37°C were grown to OD_{600}=0.8, the temperature was then lowered to 25°C and protein expression was induced by the addition of 1 mM ITPG. Next day, the culture was harvested and the protein purified as described elsewhere (Garrity et al., 2007). Pure protein was stored in 50 mM Tris-HCl (pH 7.5), 100 mM KCl and 20% glycerol. Small aliquots were immediately flash frozen in liquid nitrogen and stored at -80°C until needed.

4.6. Analytical methods

4.6.1. Agarose gel electrophoresis

For a 1% gel, 1 g of agarose was added into 100 ml of TAE buffer and heated until the agar was completely dissolved. The solution was cooled to tepid warmth before casting the gel into a chamber (BioRad) and ethidium bromide was added. The polymerized gel was transferred into a running chamber (Biorad) flooded with TAE buffer. DNA samples were mixed with an appropriate amount of loading buffer and loaded side by side with a suitable DNA ladder onto the gel. Migration of the samples was carried out at 15V / cm until band separation was completed as indicated by color separation included in the loading buffer dye. Results were examined by gel documentation using a regular camera system. DNA bands were cut using fresh cover slip glasses (20x20 mm) to avoid cross contamination.

4.6.2. SDS-Polyacrylamid gel-electrophoresis

Sodium dodecyl sulfate polyacrylamide electrophoresis was performed following the method of U.K. Laemmli (Laemmli, 1970). SDS gels were casted following standard recipes and procedures. The casted 15 % SDS gels were either used immediately or stored in a moist environment at 4°C until use.

Protein samples were mixed with an appropriate amount of SDS-loading dye and heated at 95°C for 5 minutes prior to loading into the sample wells. A molecular weight standard mix was loaded for comparison. Electrophoresis was carried out in SDS running buffer applying 60 mA until the blue dye front reaches the lower edge of the gel. The gel was stained for 15 minutes with instant blue and destained with water for 3 hours prior to documentation.

Table 14: Recipe for 15% SDS gel

Component	Stacking gel solution	Separating gel solution
4× Stacking Gel Buffer [mL]	7.5	---
4× Separating Gel Buffer [mL]	---	15
30% acrylamide (37.5:1) [mL]	4.5	30
10% APS [µL]	300	500
TEMED [µL]	30	50
MilliQ water [ml]	17.5	14.5

4.6.3. Determination of DNA and protein concentration

Protein as well as DNA concentrations were determined by photometrical methods using NanoDrop (Thermo) with implemented software controlled parameters for determination of nucleic acid - and protein concentration, respectively. Prior to sample measurement a blank measurement was carried out using the solubilizing solution only. Protein concentration was determined with respect to the individual molecular weight and extinction coefficients (Wilkins et al., 1999), while DNA was measured at absolute values assuming 1 OD corresponds to a concentration of 50 µg/ml.

4.6.4. PA5507 enzyme activity test

A previously reported coupled assay was customized to test PA5507 upon its activity using different amide substrates (Smith et al., 2009). The assay was carried out in 20 mM PO_4 buffer, pH 7.3 using 200 µM NADH, 3.3 mM alpha-ketoglutarate, 1.5 µM L-glutamic dehydrogenase (EC 1.4.1.3, Sigma-Aldrich). All substances were ordered from Sigma-Aldrich and freshly dissolved in water or 50% DMSO:water v/v. The reaction was initiated by the addition of the amide substrate in concentration of 400 µM in the presence of 1 µM PA5507. The absorption at 340 nm (ε340 = 6220 $M^{-1}cm^{-1}$) was monitored for 600 seconds at 20°C. Data was analyzed and plotted using Grafit5.

4.6.5. SAXS data collection

Small angle x-ray scattering was carried out using PA5507 buffered in 50 mM HEPES pH 8.0, 150 mM NaCl at concentrations of 6, 3, 1.5 and 0.75 mg/ml. Data were collected on a Pilatus2M at the EMBL BioSAXS beamline P12 at the PETRA-III storage ring (DESY, Hamburg). Data were processed with the ATSAS suite using PRIMUS (Konarev et al., 2003). Data analysis and comparison to the x-ray structure of PA5507 were carried out with CRYSOL (Svergun et al., 1995) and SUPCOMB20 (Kozin & Svergun, 2001).

4.7. Genetic manipulation of *Pseudomonas aeruginosa PAO1*

4.7.1. Operon selection

Other than eukaryotic genomes, bacterial genomes are, to a large extend clustered in operons where functionally linked genes are co-expressed. These operons can be predicted with confidence using bioinformatics. In this study the Database for prOkaryotic OpeRons (DOOR) (Mao et al., 2009, 2014) was used to identify potentially interesting operons. The operons chosen in this study are shown in table 16, while the selection criteria for the operons are depicted in table 15.

Table 15: Operon selection criteria in detail

Operon size	up to 4 kbp
Number of clustered genes	4 to 6
Individual protein size	15 to 90 kDa
Annotation	mostly unknown
Localization	Cytoplasmic or periplasmic
PHYRE Score	below 50
Number of homologues	low

Operon prediction was carried out by DOOR (Mao et al., 2009). PHYRE used for structural prediction resulted in scores. These refer a confidence level ranging from 0 to 100 (Kelley & Sternberg, 2009), while scores below 75 usually indicate that prediction has failed. Underlined operons were genetically knocked out in *Pseudomonas aeruginosa* PAO1 and

the ones in ***italic bold*** have been structurally characterized in this study. Structural depositions to the pdb (Berman et al., 2000) by other groups are marked in the 'notes' column.

Table 16: Operons selected for this study.

DOOR size	Gene identifier	Number of residues	Predicited function	Localization	PYHRE score	Notes
3226 bp	PA1621	270	Probable hydrolase	Cytoplasm	18	
	PA1622	286	Probable hydrolase	Cytoplasm	18	
	PA1623	220	Hypothetical protein	Cytoplasm	25	pdb: 1ECJ
	PA1624	268	Hypothetical protein	unknown	9	
3793 bp	PA3904	131	hypothetical protein	unknown	10	
	PA3905	175	hypothetical protein	unknown	35	
	PA3906	127	hypothetical protein	Cytoplasm	40	
	PA3907	261	hypothetical protein	Cytoplasm	0	
	PA3908	239	hypothetical protein	Cytoplasm	0	
1958 bp	PA4642	96	hypothetical protein	Cytoplasm	15	
	PA4643	161	hypothetical protein	unknown	29	
	PA4644	157	hypothetical protein	unknown	17	
	PA4645	185	hypoxanthine-guanine phosphoribosyltransferase	Cytoplasm	28	
3579 bp	***PA5506***	285	hypothetical protein	Cytoplasm	16	
	PA5507	217	hypothetical protein	Cytoplasm	21	
	PA5508	443	hypothetical protein	Cytoplasm	23	pdb: 4HPP (Ladner et al., 2012)
	PA5509	222	hypothetical protein	Cytoplasm	38	

4.7.2. Homologous recombination and operon knock-out

Genetic modifications of *Pseudomonas aeruginosa* strain PAO1 *in vivo* were carried out as described elsewhere (Hoang et al., 1998; Qiu et al., 2008). Mutants were generated by

the excision of the target operon from the genome using a recombination method. In detail, two fragments of about 500 base pairs flanking the end of the desired operon to knock out on either side were amplified from genomic DNA of *Pseudomonas aeruginosa PAO1*. 100ng of each purified fragment was used to extend a gentamycin resistance cassette (100 ng) amplified from pPS838 (Hoang et al., 1998) by over-lap PCR using KOD polymerase, 1mM $MgSO_4$ and 3% DMSO. The PCR protocol is shown in table 17. After the first 6 cycles, the program was paused and the gene specific outside primers were added before PCR was continued for another 25 cycles (Choi & Schweizer, 2005).

Table 17: Typical protocol for overlap PCR

Protocol	Temperature	Time	
Prepare PCR mix (25 µl): 100 ng 5'-shoulder 100 ng 3'-shoulder 100 ng Gm^{res} 200 µM dNTP mix 1 x KOD reaction buffer (5x) 0.5 µl KOD polymerase (1U/µl) 1 mM $MgSO_4$ 3% DMSO final concentration Add PCR-grade water to 25 µl	on ice		
PCR-Protocol			
Initial denaturing	94°C	2 min	
Cycle denaturing	94°C	15 sec	
Annealing	45°C	30 sec	**6x**
Extention	68°C	2:10 min	
Primer addition and			
Cycle denaturing	94°C	15 sec	
Annealing	59°C	30 sec	
		25x	
Extention	68°C	2:10 min	
Final extention	68°C	5 min	
hold	4°C		

The generated fragments were inserted into linearized plasmid pEX18Tc at the *SmaI* restriction site. Its non-*Pseudomonas* origin of replication drives genomic integration. The plasmid also contains the *sacB* gene from *Bacillus subtilis* which encodes for exoenzyme levansucrase (Gay et al., 1983) providing a marker for negative counterselection during the recombination process, which has been already reported to be successful in many cases (Gay et al., 1985; Pierce et al., 1992; Frengen et al., 1999; Blommel et al., 2009).

Table 18: Example of 3‘- and 5‘-shoulders generated and used in overlap PCR with gentamycin cassette

PCR fragment	1621m1_2	1624m3_4	Gm^r	1621m1_2 + Gm^r	1624m3_4 + Gm^r
Size (bp)	680	599	1057	1720	1639

The furnished plasmid was transformed into electro-competent *Pseudomonas aeruginosa* as described elsewhere (Choi et al., 2006). Cells that integrated the plasmid into their genome using the provided 5’-end recombination site were selected on agar containing 300 µg/ml gentamycin. Next, the second step of recombination was triggered by plating selected clones on both 5% sucrose containing agar as well as on gentamycin spiked agar. Colonies that lost *sacB* grew under both conditions. Colonies were patched on new plates, incubated overnight and glycerol stocks prepared. Target operons were replaced by the gentamycin resistance cassette in selected bacterial colonies. All cloning and recombination steps were checked by PCR using specific targeting primer (Poteete et al., 2006) and verified by DNA sequencing. A full list of used oligonucleotides is shown in the appendix.

4.7.3. Extraction protocol for metabolomic analysis

Pseudomonas aeruginosa mutants were grown in 60 ml Miller-Hinton II media at 37°C, shaking at 140 rpm. The broth was inoculated with a single colony, picked from double streaked agar plate. The protocol is separated into two parts. First, the dry cell weight was determined by using the mean of three technical replicates. Secondly, the metabolite extraction was carried out using three biological replicates. All solutions were purchased in minimum purity grade of ‘HPLC-MS’ (>99.95%). Working quickly but at the same time precise was indispensable during the process of metabolite extraction.

Standard HPLC glass vials were dried to constant weight at 60°C for a minimum of 3 hours before weighting them (vial weight). After three hours, 10 ml of culture volume was collected, its optical density noted (ideal was 2.0) and cell harvested in a pre-cooled centrifuge at 4°C for 5 min at maximum speed. The supernatant was discarded and

residual media carefully removed, not to disturb the pellet. Cells were twice washed with water, harvested and the supernatant discarded. The resulting pellet was resuspended in 1 ml of water, transferred to the glass vial and stored in a drying furnace until constant weight and dryness before the new weight is determined. The dry cell weight (DCW) for this specific optical density is calculated by subtracting the vial weight from the complete weight.

Metabolite extraction was carried out with a DCW of 10 mg. The necessary culture volume (at OD_{600} = 2) was poured into a 50 ml falcon side by side with 10 ml fresh media serving as a blank sample. Both samples were treated to same way. In parallel the same volume was used to determine the dry cell weight, as described above. The filled falcon was drown into liquid nitrogen and inverted before it was dropped back for 5 seconds, 5-6 times until the culture gets to ~10° C. Appropriate filter (PALL, TF-450, PTFE, lot#: 40813) were equilibrated to saline conditions. The cooled culture was poured into the -20° C cold vacuum filtering manifold equipped with equilibrated filter. The filter was twice washed with 5 ml ice-cold saline and then quickly transferred into a sterile 50 ml falcon tube and placed in a dry-ice/ethanol bath. The extraction was carried out in the bath by pipetting 1 ml extraction buffer (80% MeOH in water, -20°C) several times over the filter to detach cells from the filter. Last, the filter was removed and the flacon immediately transferred into liquid nitrogen.

4.8. Principles of Crystallography

4.8.1. Concept

In biochemical science knowledge about a spatial macromolecular structure is of importance to understand the function of a protein on the molecular level. Protein-protein complex structures have been determined by cryo electron microscopy (Cryo-EM) to moderate resolution (Vinothkumar et al., 2014) and nuclear magnetic resonance (NMR) is employed to gain structural information of proteins in solution (Mineev et al., 2014). Yet, x-ray crystallography is the only technique that reliably provides the answer to structural questions from global fold to atomic resolution of chemical bonds (Howard et al., 2004; Wang et al., 2007; Mueller et al., 2011).

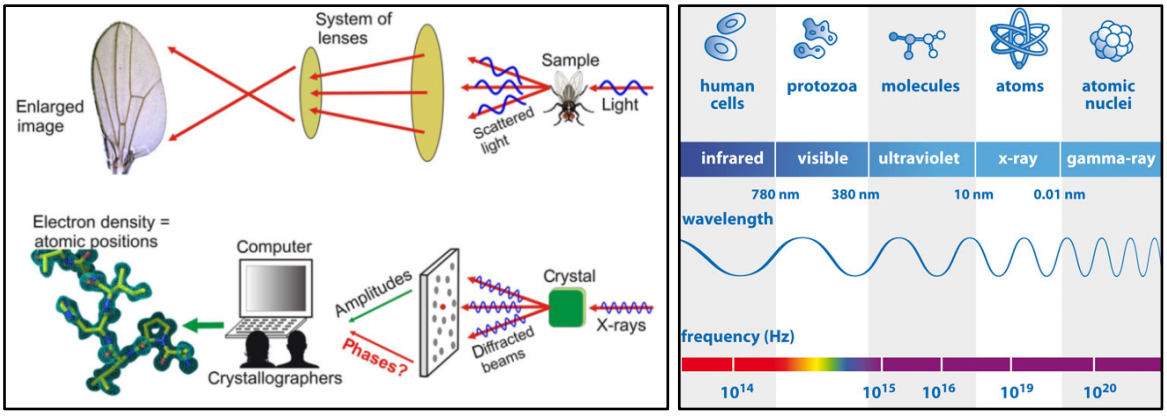


Figure 13: Comparison of light microscope depending on lenses and x-ray crystallography, adopted from (Miyaguchi, 2014) on the left. X-rays are part of the electromagnetic spectrum with λ between 10 and 0.01 nm.

Resolution (d) is the capacity of an optical system to resolve two objects as separate images and is defined by the wavelength (λ) of the employed light and the refractive index (n). This dimensionless parameter is given by the quotient of the velocity of light in vacuum (c_v) and in a given material (c_m). Aqueous protein solutions and crystals thereof show a faintly higher refractive index than water itself (1.33), ranging from 1.332 through 1.45 (Cole et al., 1995).

$$d = \frac{\lambda}{2 * n} \qquad with \qquad n = \frac{c_v}{c_m}$$

Equation: Resolution limit (d) is defined by the wavelength λ and the refractive index n. Reformulation of the Bragg equation defines resolution with sin θ scattering angle of the ray collected on the detector

Light microscopy has its maximum resolution at 0.25 µm, justified by the light used (λ = 400 – 700 nm) (Goodwin, 2014). This does not match the requirements to enlighten atomic details of a protein structure by orders of magnitude. Assuming two carbon atoms with an average distance of 0.15 nm need to be resolved, the wavelength of the light used must be around 0.3 nm, falling to the range of X-rays (figure 13). The need of these high energetic rays raises two inherent issues. First, other than visible light, x-rays cannot be refracted by any lens system. Secondly, single protein molecules refract a x-ray beam weakly; therefore a protein crystal is necessary with coherently arranged molecules in a defined crystal lattice utilizing the concept of constructive inference. Scattering information generated by every molecule arranged in the crystal contributes to quantitative diffraction and can only be detected and recorded in cases of deflected constructively interfering scattered rays.

4.8.2. Protein crystals

Protein crystals contain on average 50% solvent, which renders them rather fragile compared to a salt crystal (Matthews, 1968; Chruszcz et al., 2008). As a consequence and in contrast, macromolecular crystals need to be grown from aqueous solution, explaining their fragility and disintegration upon dehydration and sensitivity to all kinds of environmental variations.

The physical background of crystal formation is not yet fully understood (Haas & Drenth, 1999). Considering a phase diagram (figure 14) with a metastable immiscibility region, crystallization of proteins from a solution follows a distinct two-step process with two inseparable steps: nucleation and crystal growth. The process involves a transition at which protein molecules are present in a concentration higher than their solubility level. Stabilized by the physical and chemical environment of the mother liquor macromolecules in these metastable liquid droplets allocate from fully disorder into an ordered state in para-crystalline nuclei (McPherson et al., 2000). With the momentum of the reestablishment of the protein equilibrium nuclei spread out and become larger by the attachment of molecules in all dimensions, leading to three-dimensional nucleation (McPherson et al., 1995, 2000). The local concentration is lowered and drops back into the metastable phase where further crystal growth takes place.

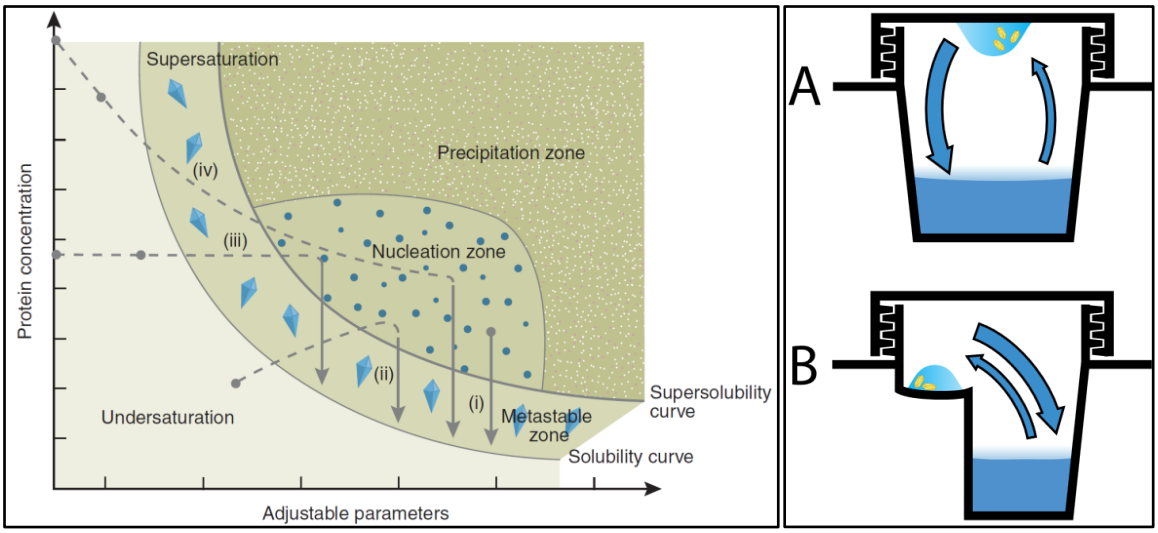


Figure 14: Left: Schematic phase diagram: Nucleation occurs in the supersaturated labile phase, while crystal growth happens in the metastable supersaturated phase, adapted from (Chayen & Saridakis, 2008). Right: Setup of a hanging drop (A) and sitting drop (B) vapor diffusion crystallization experiment, adapted from Wikipedia.

In theory the supersaturated state can be reached by the change of various parameters, e.g. temperature, salt concentration, addition of polymers and/or precipitant concentration, in practice, many parameters are tuned in a controlled way at the same time. Initial crystallization conditions are mostly searched with commercial screening suites. Once drops containing undersaturated protein solution and premixed mother liquor are set up the compartment is tightly sealed. The restricted correspondence only between the small protein containing drop and the mother liquor reservoir attracts water molecules by the slightly higher concentration of precipitant to the reservoir, thereby elevating the protein concentration towards supersaturation. Added Polymers induce molecular crowding by the volume-exclusion effect that induces separation of solution and protein molecules (Ingham, 1984, 1990; Englard & Seifter, 1990). Systematic optimization of hit conditions was carried out until desired crystals are isolated and cryo-protected to minimize radiation damage and extend the life time of the crystal prior to exposer to an x-ray beam at 100 Kelvin.

A diffraction pattern generated by a protein crystal is shown in figure 15. Its diffraction limit can be correlated with its degree of internal order, called mosaicity. Theoretically, the more uniform molecules are arranged with the protein lattice, the higher is the resolution of diffraction. Compared to salt crystals which diffract usually up to their theoretical diffraction limit, protein crystals are more restricted to other parameters like

unstructured bulk water molecules (Figure 15). They diffuse due to the Brownian motion (Kakutani, 1950) inside solvent channels of the protein framework and do not contribute to diffraction, moreover they limit theoretical resolution due to “diffuse scattering” (Shpyrko et al., 2004; Xu et al., 2012).

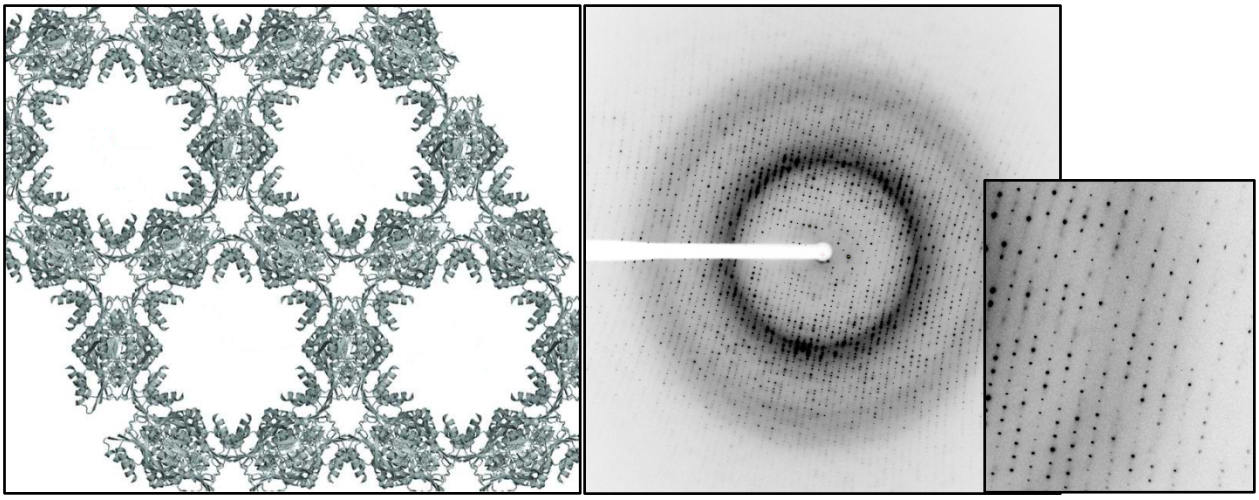


Figure 15: Left: Crystal packing with a water content of more than 60%, adapted from (Xu et al., 2012). Right: experimental x-ray diffraction patterns of a protein crystal collected on a Rayonix CCD detector with magnification.

4.8.3. Bragg’s Law

Diffraction of x-rays by crystals was first discovered by Max von Laue. W. H. Bragg and his son derived an equation by treating diffraction as reflections from crystal lattice planes.

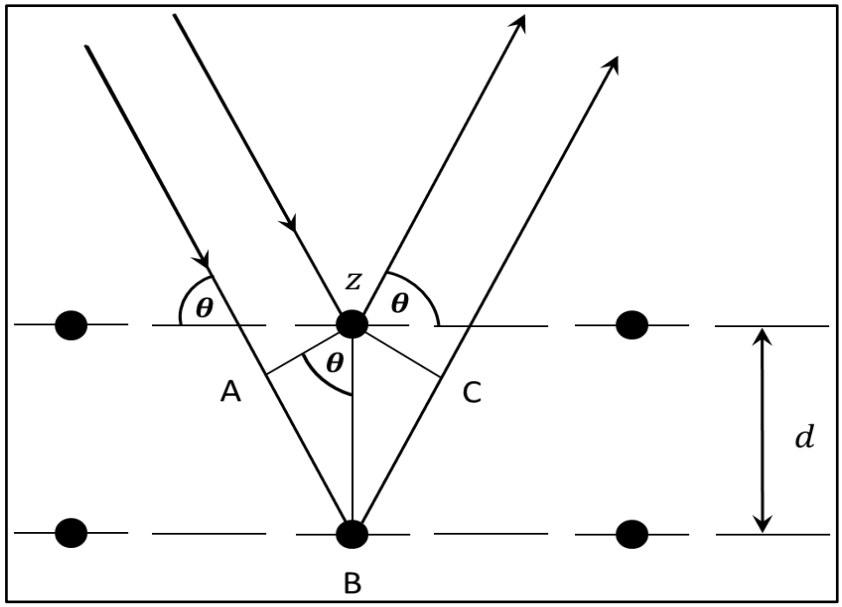


Figure 16: Deriving Bragg's Law using the reflection geometry and applying trigonometry

X-rays striking a crystal at an incident angle are scattered. In order for the waves to interfere constructively, the spacing between the lattice planes must be equal to integer multiple of the wavelength used. When this constructive interference occurs, a diffracted beam of X-rays will leave the crystal at an angle equivalent to that of the incident photon beam. Because of experimental setup rays are of discrete wavelength and in parallel phase when striking the different layers of the protein crystal. Coherent incident photons are scattered by the top layer (A), while other photons continue travelling to the next layers where it is scattered by different atoms (B). The diffracted photon at atom B must travel the extra distance AB and BC. In order to continue traveling in phase, this distance d must be an integral multiple of the wavelength λ (figure 16).

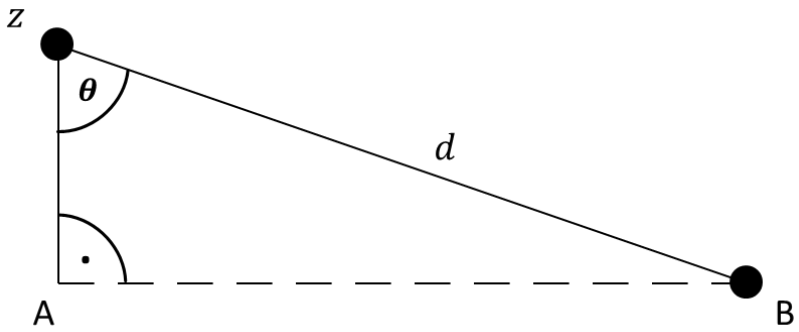


Figure 17: Extracted triangle from Bragg's geometry

Right triangle ABz extracted from figure 17 recognizes d as hypotenuse. Using trigonometry were $\sin\theta = AB/d$ results in

$$AB = d\sin\theta \qquad (2)$$

Since AB = BC as depicted in figure 8 follows

$$n\lambda = 2AB \qquad (3)$$

Substitution of AB in the equation 3 by 2 result Bragg's Law:

$$n\lambda = 2d\sin\theta \qquad (4)$$

Equation: Bragg's Law, explaining crystals reflect an X-ray beam at certain angles and manner. The variable lamba λ is the wavelength of an x-ray beam; d is the distance between lattice layers in the crystal lattice and n an integer number.

Parallel incident beams that hit the same set of lattice planes at the angle θ produce a signal in direction of the reflected beam if the Bragg's law of diffraction is satisfied.

4.8.4. Real space and reciprocal lattice

Protein molecules are organized in a crystal lattice within the crystal, the space group, a type of unit cell together with any symmetry relation with the molecules organized. This organization can be described by lattice planes which are constructed planes hitting lattice points somewhere in the lattice (figure 18). This lattice planes are independent of the space group.

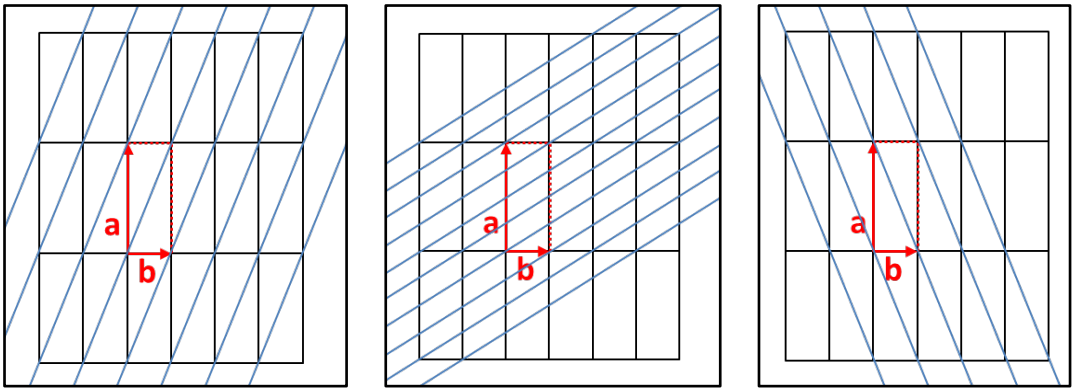


Figure 18: Crystal lattice planes of different indices: left (1 1 0) – middle (4 1 0) – right (1 -2 0)

A set of crystal planes can be uniquely identified by the three numbers h, k, l. The lattice plane (hkl) runs through origin and point a/h b/l and c/l. The integer number of h k and l defines how many times the axis a b or c is cut by the set of planes. If this index is zero, the set of lattice planes is parallel to the respective unit cell axis.

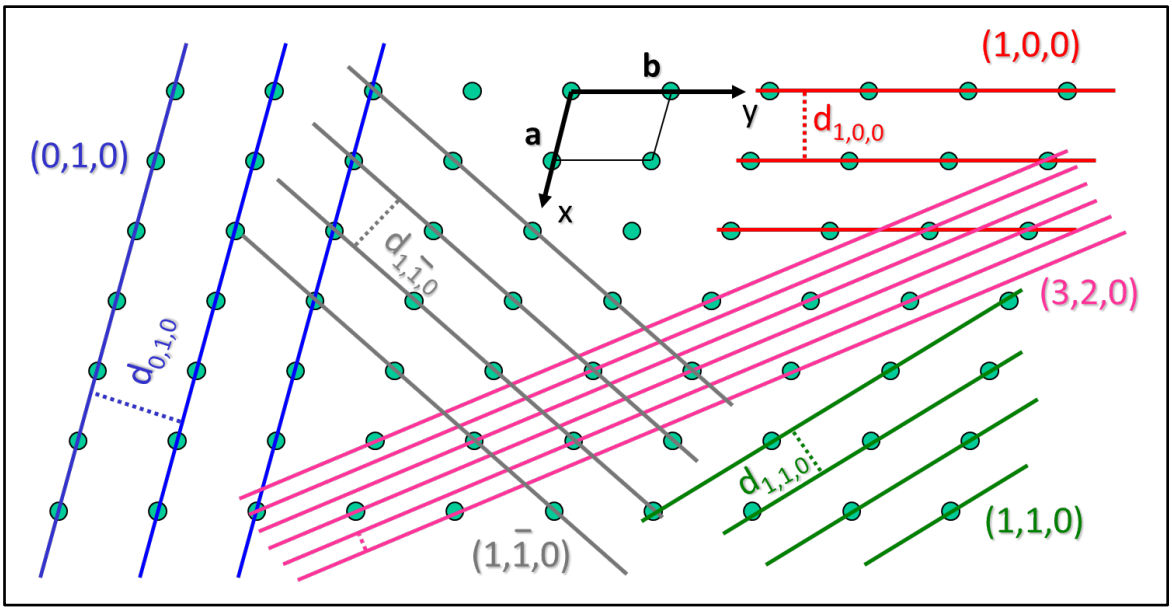


Figure 19: Miller indices of lattice planes.

Each direct lattice is connected to a reciprocal lattice. The reciprocal lattice is built from an origin O by construction of all possible normals to a direct lattice plane (hkl). They are determined at distance $1/d_{hkl}$ form the origin where d_{hkl} is the perpendicular distance between planes of a lattice plane set (hkl). All end points of these vectors produce a periodic lattice, the reciprocal lattice. The points, obtained on that way are identified by a numerical triplet, called the Miller indices (hkl) (figure 20).

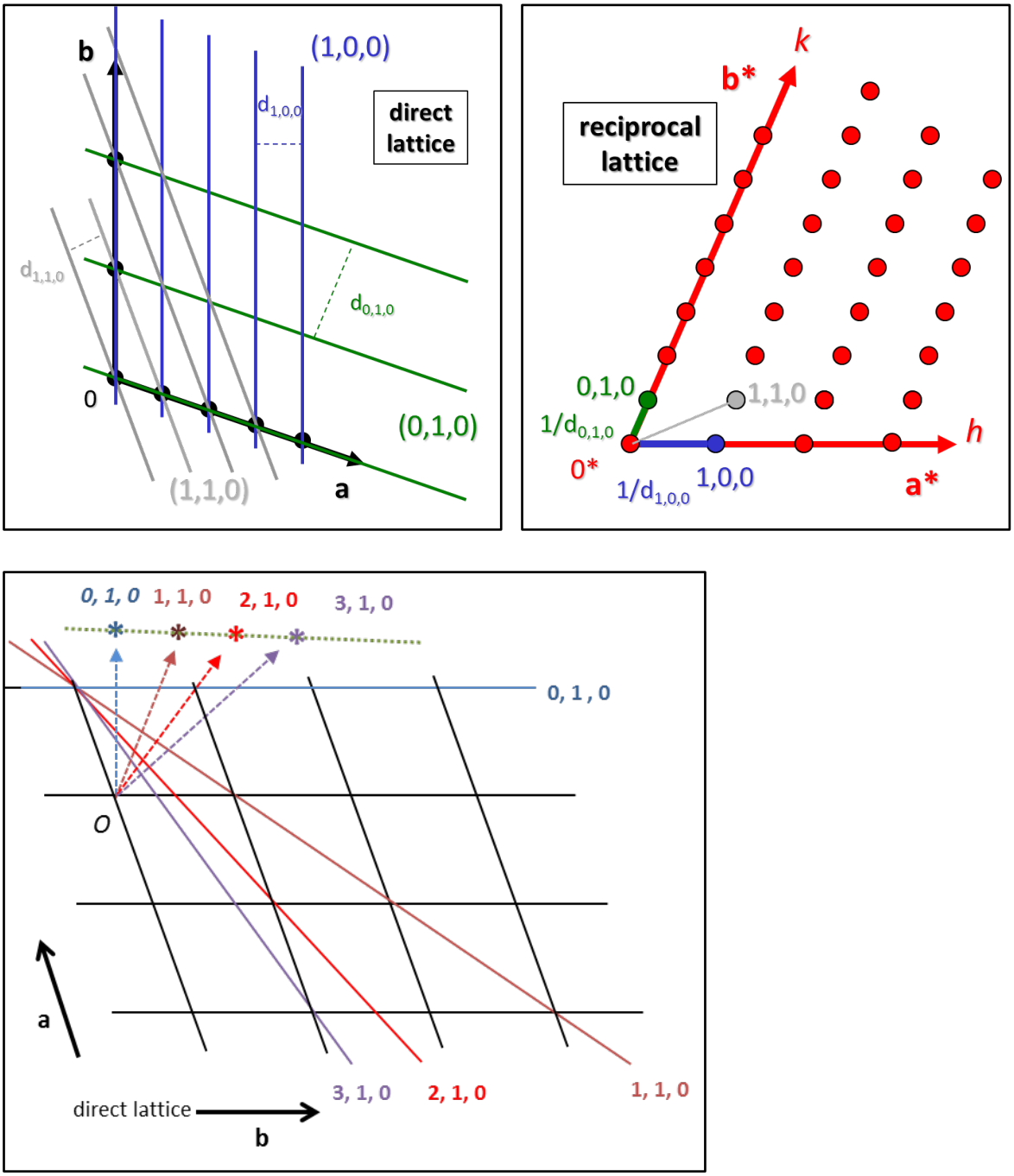


Figure 20: Geometrical construction of a reciprocal lattice point from the direct lattice. To simplify, the third axis of the direct lattice (c) is assumed to be perpendicular to the screen.

The reciprocal unit cell can be calculated, depending of the angles and the dimensions of the direct unit cell lattice. Assuming α = β = γ = 90° the reciprocal unit cell vectors $|\boldsymbol{a}^*|, |\boldsymbol{b}^*|, |\boldsymbol{c}^*|$ can be calculated as given as

$$|a^*| = {}^{1}/_{|a|} \quad a^* \textit{ parallel to } a$$
$$|b^*| = {}^{1}/_{|b|} \quad b^* \textit{ parallel to } b$$
$$|c^*| = {}^{1}/_{|c|} \quad c^* \textit{ parallel to } c$$

Equation: Calculation of reciprocal unit cell dimensions from real space dimensions.

4.8.5. Bragg's law in reciprocal space – Ewald construction

The Ewald construction represents a crystal by its reciprocal lattice points. The Ewald sphere is a three-dimensional construction with the radius $1/\lambda$. Reciprocal lattice points which coincide with the Ewald sphere are satisfying Bragg's law and are therefore contributing to constructive interference of the scattered beam. For a random orientated crystal that is rotated during an experiment, the reciprocal lattice will rotate as well (figure 21). Reciprocal lattice points get in contact with the constructed Ewald sphere and diffraction will occurs. Figure 21 illustrates this for the case where *h*,*k*,*l* = 2,0,0. The collected diffraction image is a two dimensional projecting of the three-dimensional reciprocal lattice on the Ewald sphere.

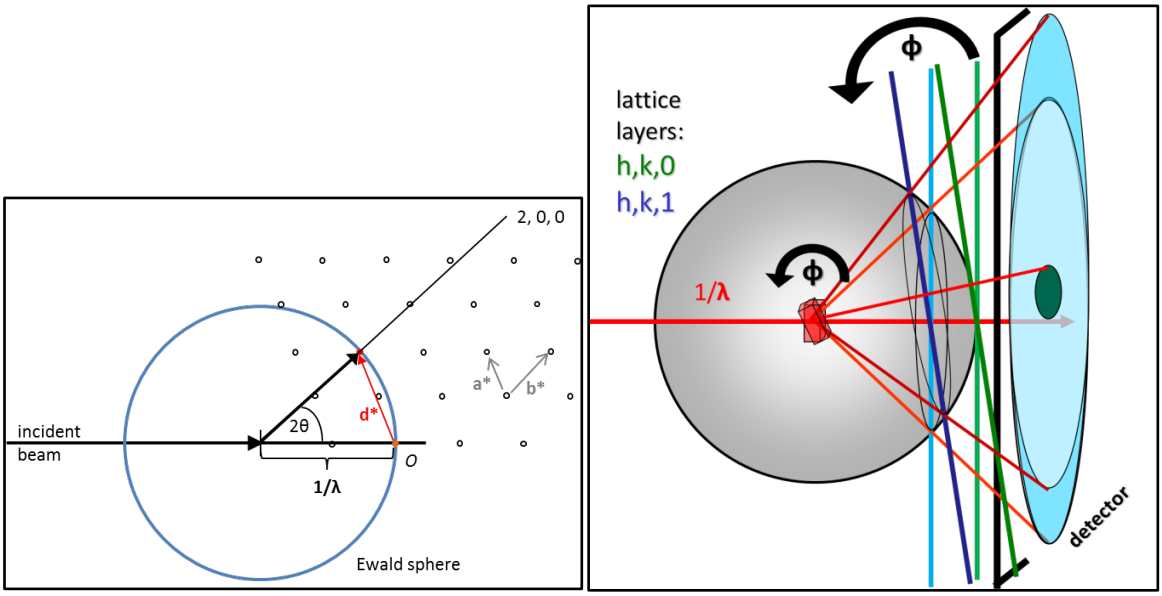


Figure 21: Ewald sphere construction: Constructive interfering diffraction occurs at reciprocal lattice points interacting with the Ewald sphere with the radius 1/λ.

4.8.6. Diffraction to electron density – Fourier transformation

The goal of every diffraction experiment is to determine the electron density *ρ(x,y,z)* for all x, y and z in the unit cell. During data collection diffraction images are recorded including information about the position (hkl) and the intensity (I_{hkl}) of each reflection within a diffraction pattern. The structure factors can be calculated if the content of the unit cell in known. Thus, crystallographic experiments deal with the inverse problem. Information regarding the structure factors is available but the content of the crystal is unknown. If the structure factor equation is applied to a Fourier transformation the equation describing the electron density in the crystal is yielded. The Fourier transformation of the diffraction data is the representation of the crystals content, the organization of macromolecules. And a Fourier transformation of the content of the crystal results in the diffraction pattern. This correlation is shown in the following figures is detail.

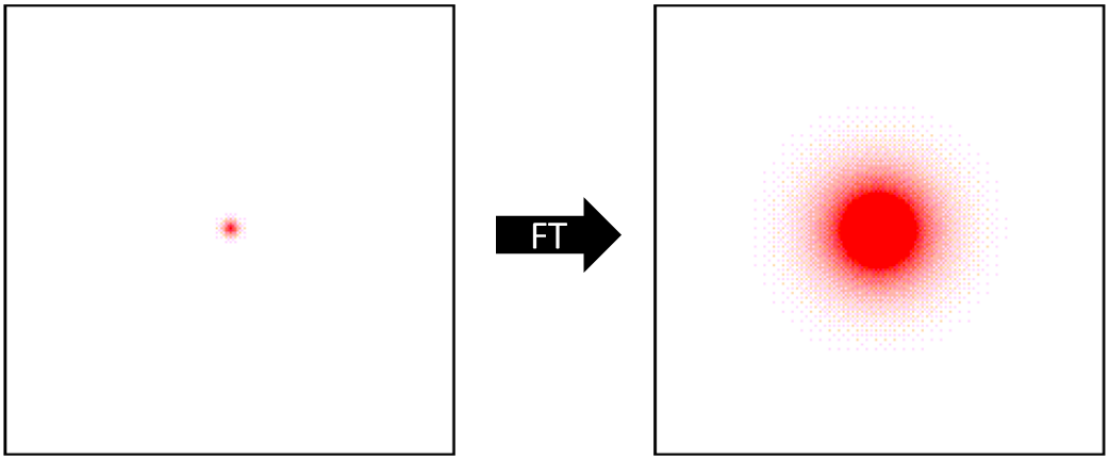


Figure 22: An atom and its Fourier transformation. The sharp atom of circular symmetry results in a broad smooth function featuring also circular shape. FT = Fourier Transformation.

The precise shape of the single atom and its Fourier transformation do feature the same shape, a smooth circle. The transformation of a more complex molecule is depicted in the figure 22.

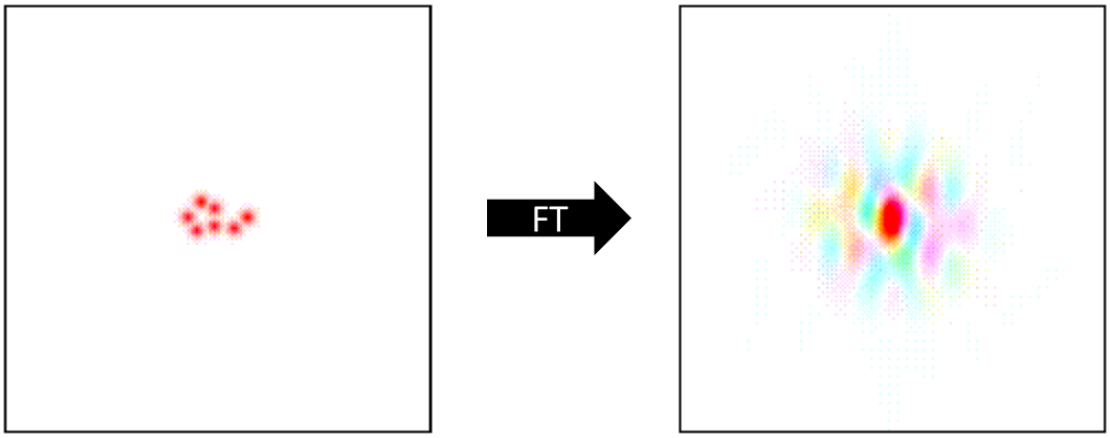


Figure 23: Fourier transformation of a more complex molecule shows some details, but the overall shape is not the in direct relation to the molecule transformed.

Now taking a crystal into account were molecules are arranged into a lattice, shown as grid point in figure 23. The transformation of this grid results in a new grid with reciprocal directions and reciprocal distances between them.

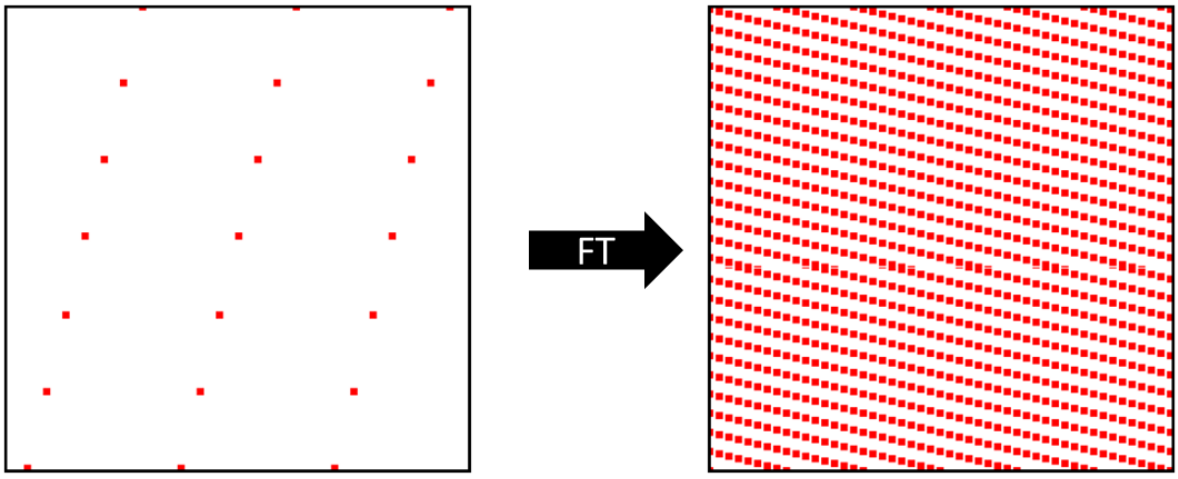


Figure 24: Transformation of a real space lattice (left) results in the reciprocal lattice shown on the right side. The distances and x,y,2θ coordinates within the transformed lattice are reciprocal to the real ones.

In a protein crystal molecules cannot be considered as single atoms of grid points but are evenly assembled macromolecular molecules arranged in a crystal lattice. In a crystal the more complex molecules (figure 23) are arranged in a grid (figure 24). The resulting Fourier transformation is therefore the combination of the transformation of the molecules with the reciprocal lattice, resulting in the diffraction pattern (figure 25). Every reflection spot does consequently contain information about all atoms in the crystal.

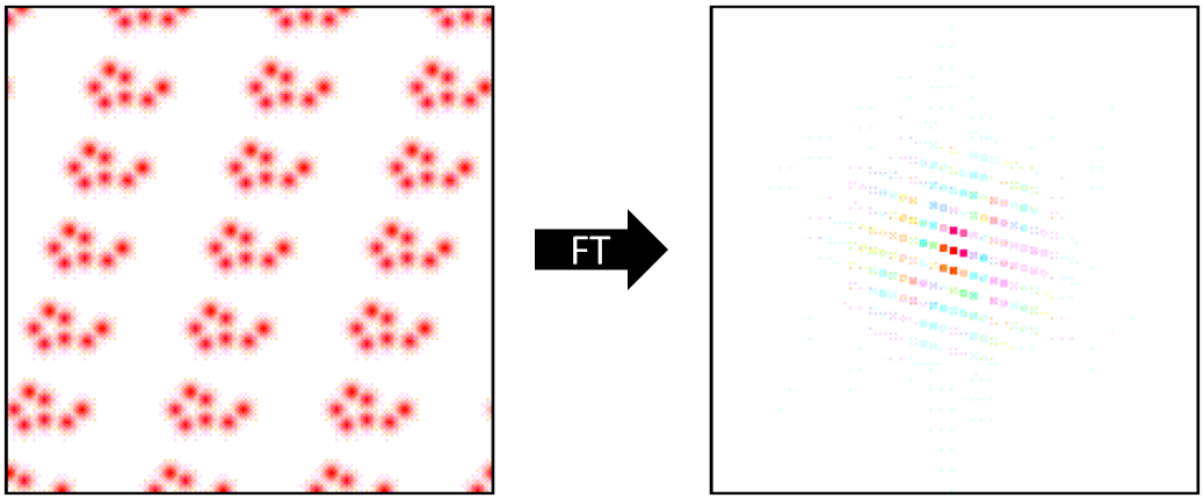


Figure 25: Fourier transformation includes molecular information spread on the reciprocal lattice of the crystal.

During data collection acquired images contain reflections of an x-ray beam which was diffracted on the electron envelope of a protein arranged in the unit cell. In principle the electron distribution in the crystal is depicted within the images, however, a direct interpretation and reconstitution is not possible because the collected data does only contain geometric position (index hkl) and attributed intensities (I_{hkl}) for each reflection. The diffraction pattern is related to the object scattering them. The mathematical operation needed to convert the pattern in a three dimensional electron density is called Fourier transformation. If the electron density is the mathematical function, then the diffraction pattern is the Fourier transform of that function. A Fourier transformation in either direction requires structure factors consisting of two things: amplitude and phase. During data collection the information containing the phase is lost, which results in the phase problem. Since there is no practical way to measure relative phase angles for the different reflections experimentally, the only parameter that is measured amplitudes and the intensity of each spot on the detector called $|F_{obs}|$.

The structure factor $\overrightarrow{F_K}(hkl)$ includes the length of the vector but also the phase $\alpha(hkl)$.

$$\overrightarrow{F_K}(hkl) = |F_K(hkl)| \cdot e^{i\,\alpha(hkl)}$$

The electron density is calculated by a Fourier transformation of all structure factors $\overrightarrow{F_K}(hkl)$.

$$\rho(xyz) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}\overrightarrow{F_K}(hkl)\cdot\ e^{-2\pi i(hx+ky+lz)+i\varphi_{hkl}}$$

The phases can be derived experimentally from a scattering experiment (α_{obs}) or they can be borrowed from an existing similar structure during a molecular replacement solution (α_{cal}).

The importance of the phase is shown in figure 26, where a phase error is shown to introduce model bias.

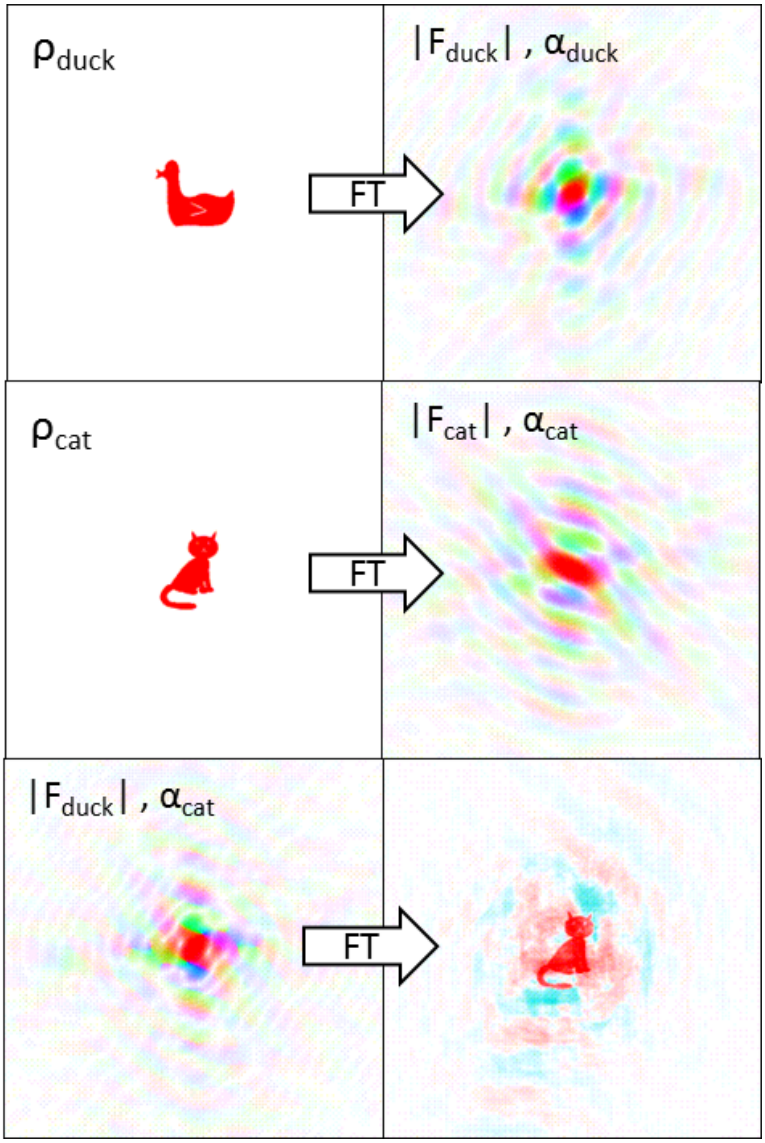


Figure 26: Model bias introduced by wrong phasing

4.8.7. Molecular replacement and experimental phasing

As described previously the phase problem can be solved by either molecular replacement, which is restricted by the existence of a similar protein structure used as search model or experimental phasing by anomalous diffraction.

Molecular replacement uses the phase from an existing model structure and combines this with the experimentally measured amplitudes of a target protein structure, depicted in figure 27. In practice this method, however, can only be employed if the target protein shows high similarity to the already existing model. If the similarity is not given, model bias is introduced and the resulting solution does not fit the electron density that usually is discontinues itself.

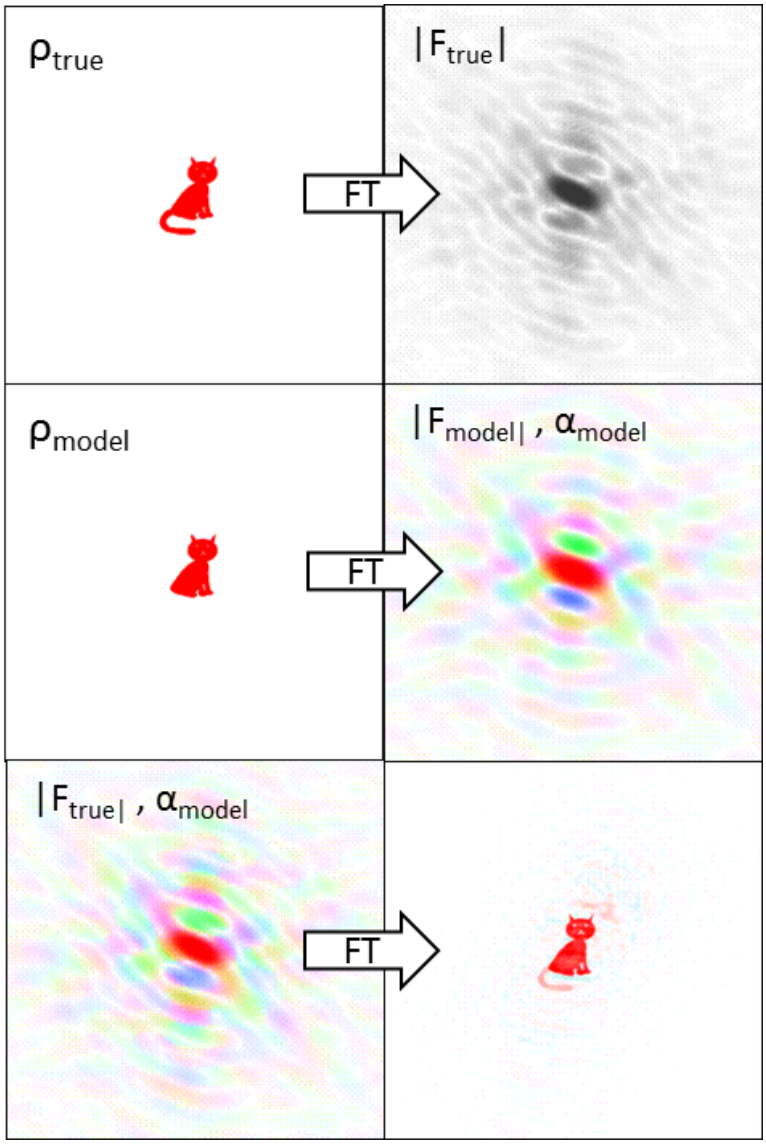


Figure 27: Molecular replacement scheme: The phase of the tail-less model-cat is used to phase the amplitude of the cat.

Experimental phasing is used by introduction of anomalous scattering atoms in the structure. Data at and near the absorption edge of the incorporated anomalous atom is collected. Heavy atom positions are located in calculated difference Patterson maps from which phases can be calculated. During normal scattering events incident photons have energies not matching the electronic transition of atoms in the crystal, resulting in elastic Thomson scattering by outer electrons (figure 28). The scattering is described by the atomic scattering factor $\mathbf{f^o}$.

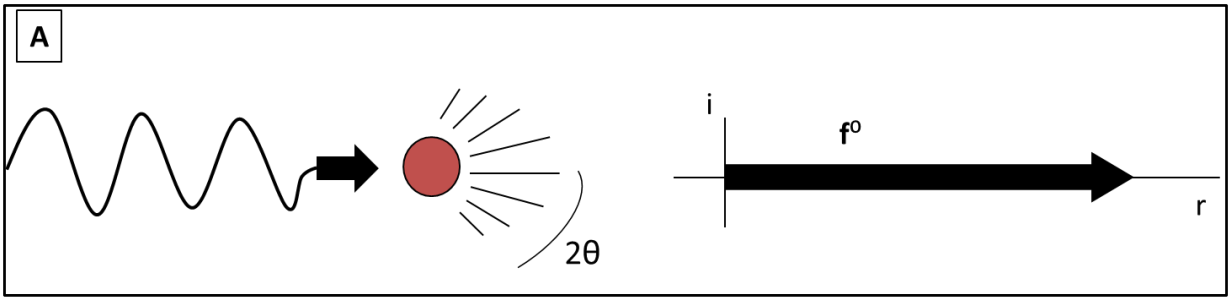


Figure 28: Event of „normal scattering“: The scattering factor f^o is purely real and independent of the wavelength

Incident photons will scatter in case of insufficient energy responsible for causing a transition of an electron. However, if incident x-rays approach or match the energy of an element, electronic transition occurs and x-rays are absorbed and only part of the photon's energy is scattered. This is called anomalous scattering. The reemission of the absorbed energy is possible either by fluorescence, emitting light at lower energy, which is used during an x-ray fluorescence scan; or it can be emitted at the same energy as the incident photon but involves perturbation of the scattered photon. The event involves a wavelength dependent phase shift which is independent of the scattering angle θ due to its origin in bound core shell electrons as depicted in figure 29.

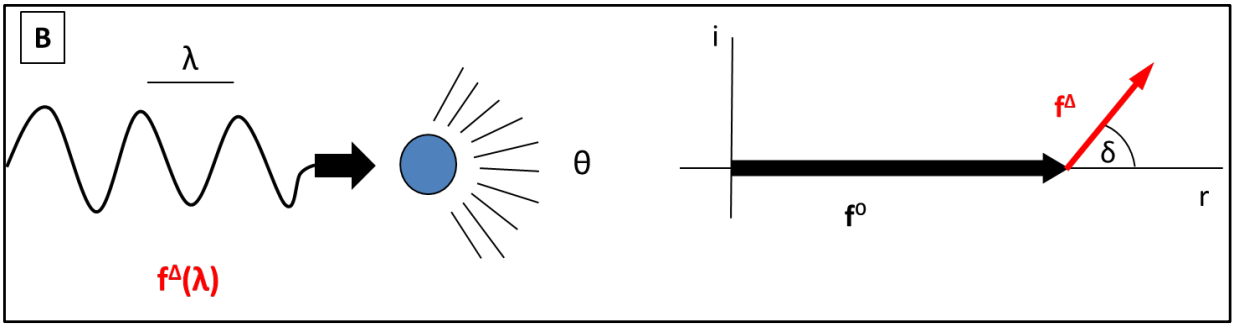


Figure 29: Anomalous scattering event.

In greater detail the anomalous scattering vector as depicted in figure 18 contains an imaginary as well as a real part. Figure 30 elucidates the scattering event in greater detail.

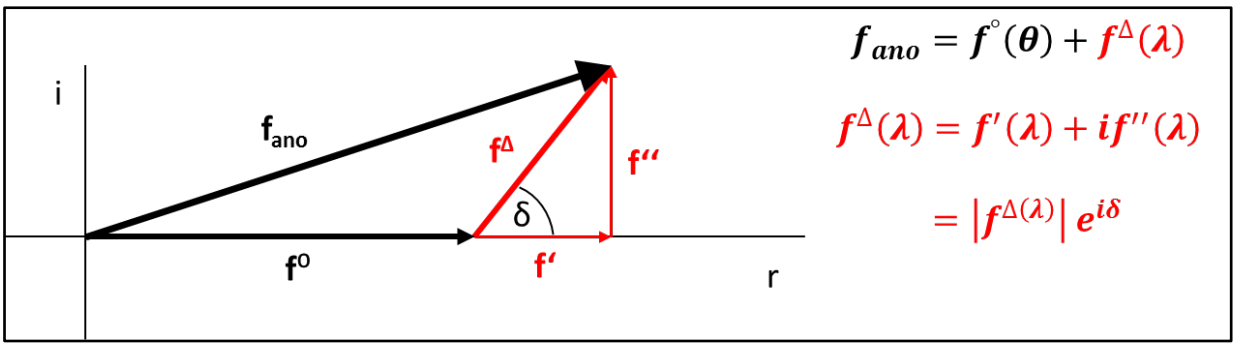


Figure 30: Atomic scattering factor of an anomalous scatterer f_{ano}.

The anomalous scattering factor f_{ano} is calculated from a real part f' and the imaginary f'' as followed:

$$f_{ano} = f^{\circ}(\theta) + f^{\Delta}(\lambda) = f^{\circ}(\theta) + |f^{\Delta(\lambda)}|e^{i\delta} = f^{\circ}(\theta) + f'(\lambda) + if''(\lambda)$$

The f° parameter is purely real and independent of wavelength. The scattered intensity falls off with increasing scattering angle (θ) due to diffuseness of the outer electron shells of the scattering atoms (Hendrickson, 1991). The second $f^{\Delta}(\lambda)$ is complex with the real $f'(\lambda)$ and the imaginary part $if''(\lambda)$. In contrast to f°, f^{Δ} is depended on a discrete wavelength but virtually independent of scattering angle. The parameters (f' and f'') are directly determined from an adsorption – fluorescence scan prior to accurate anomalous data collection on at a tunable beamline with high brilliance. High redundancies are mandatory because the anomalous signal is small and adds up only during data collection.

4.8.8. Physical background – anomalous scattering

Friedel pairs are Bragg reflections related by inversion through the origin and according to Friedel's law they have equal amplitude but opposite phase.

$$|F_{hkl}| = |\overline{F_{hkl}}| = |F_{-h-k-l}| \quad with \quad \varphi_{hkl} = -\overline{\varphi_{hkl}} = -\varphi_{-h-k-l}$$

The diffraction pattern of a crystal is centro-symmetrical which relates pairs of the same but opposite Bragg reflections in a given space group symmetry – the Bijvoet pairs F^{+}, F^{-}.

In the case of anomalous scattering Friedel pairs and subsequently Bijvoet pairs do not match anymore so that $|F_{hkl}| \neq |\overline{F_{hkl}}|$ and $\varphi_{hkl} \neq -\overline{\varphi_{hkl}}$ as well as $|F^+| \neq |F^-|$ resulting in the anomalous difference given by $|\Delta F_{ano}| = |F^+| - |F^-|$. This relation is shown graphically in figure 31.

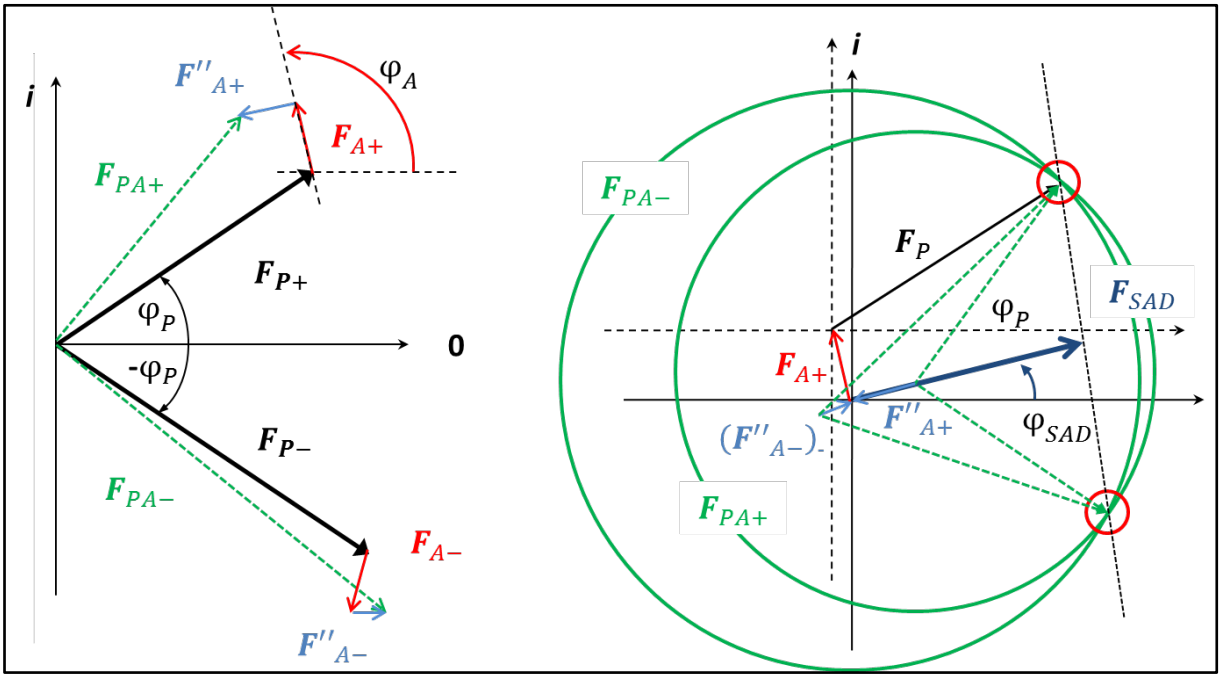


Figure 31: Anomalous phase information form SAD data – breakup of Friedel's Law. The protein structure factors (F_{P+}) and the Friedel mate (F_{P-}) and the contribution of the anomalous scatterer. Figure adopted from 2011 From Biomolecular Crystallograph by Bernhard Rupp.

4.8.9. Model building, refinement and structure validation

In case of molecular replacement, coordinates were put in phenix.refine (Afonine et al., 2012) for initial refinement. The output model was used as starting model for iterative rounds of manual adjustments by COOT and further phase improvements and refinement by phenix application refine. Structural error identification and validation was done by Molprobity (Chen et al., 2010).

Initial model resulting from either SHELXE (Sheldrick, 2010) or AutoSol (Zwart et al., 2008; Adams et al., 2010) were corrected manually in COOT (Emsley et al., 2010) and the Phenix application Refine (Afonine et al., 2012) was used for the automated refinement throughout the model building process. Water molecules were introduced manually and structures validated by Molprobity (Chen et al., 2010).

Determined phases may contain errors which are reflected in errors in the electron density $\rho(x\,y\,z)$. These errors are corrected by structure refinement, both by manual adjustments in real space using visualizing programs that allow structure manipulation and in reciprocal space carried out by computer software. Knowledge about the building blocks of the model and its composition allows modifications to adjust the model as much as possible to the experimentally determined electron density. The refinement is carried out in reciprocal space. First, structure factors of a model are calculated by Fourier transformation, $|F_C|$. Second, the experimentally determined structure factors are compared to model one by using sophisticated algorithms (McCoy, 2002; McCoy et al., 2004; Pannu & Read, 2004) which are trying to minimize the difference between both by optimizing parameters like atom positions, B-factors, stereo-chemical restrains, etc. All these parameters are mapped into a function that is further optimized in a way that the function is minimized.

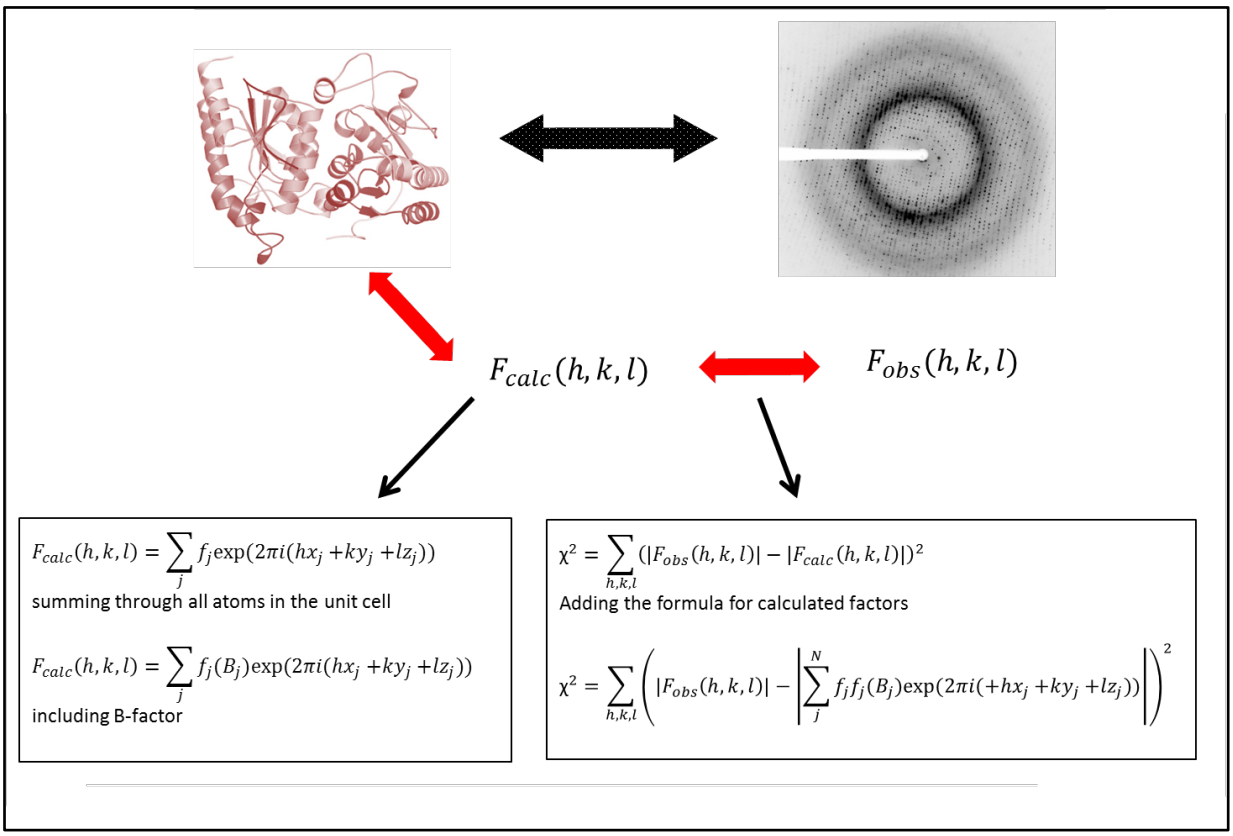


Figure 32: Initial refinement formula for refinement

An overview of collected heavy atom data is given in table 19. Full data collection statistics are summarized in the appendix (table 43).

Table19: Datasets used for structure determination

Project	Dataset	λ(Å)	Space group	Software
PA1622	**SeMet_pk**	0.9760	$P6_1$	SHELX
PA1623	**SeMet_pk**		$P2_12_12$	Phenix.AutoSol
PA1624	**SeMet_pk**	0.9800	$P2_12_12_1$	Phenix.AutoSol
PA5506	**SeMet_ip**	0.97992	$P2_12_12_1$	SHELX / Phenix.AutoSol
	SeMet_pk	0.97981	$P2_12_12_1$	SHELX / Phenix.AutoSol
PA5506	**SeMet_pk**	0.979531	$P4_12_12$	SHELX / Phenix.AutoSol

4.9. X-ray crystallography methods

4.9.1. Crystallization

Initial crystal screening was performed by the sitting drop vapor diffusion method using JCSG Core Suites I – IV, the Cryos (Qiagen), MidasSuite and Morpheus Screen (Molecular Dimensions). Either native or labeled purified protein solution was used at three different concentrations. Drops consisting of 200 nl and 200 nl mother liquor were prepared with nanodispensing robot (Phenix, Art Robbins Instruments or HoneyBee, Digilab Genomic Solutions) on an MRC3-drop plate (Molecular Dimensions) or Intelli Plate (Art Robinson). All crystallization trails were incubated at 20°C in a temperature controlled imaging system (Formulatrix). Initial crystals hits were further optimized in size and shape during several rounds of grid screening optimization with 1 µl protein plus 1 µl mother liquor using hanging drop plates (Linbro), equilibrated against 500 µl of mother liquorFinal reservoir compositions are summarized in table 44 (appendix)

Crystals were recovered from mother liquor and transferred into a new drop containing mother liquor supplied with a cryo-protecting agent, listed in table 44, prior to plunging into liquid nitrogen.

4.9.2. Data collection

Crystallographic data were collected at 100 K either in house on a sealed tube (Incoatec) with a MAR345 image plate detector or a Rigaku MicroMax-007 HF generator with a Rigaku Saturn 944+ CCD detector; both systems were equipped with an OxfordCryo stream system (Oxford Cryo 700). High-resolution data and anomalous data were collected on the X10SA beamline at the SLS, Villingen, Switzerland, the beamline BL14.1 and 14.2 at BESSY II, Berlin (Mueller et al., 2012) as well as on beamline ID23-1 (Nurizzo et al., 2006a) at ESRF, Grenoble, France. Diffraction was recorded on either a Rayonics MX-225 CCD using fast shutter control or on a Pilatus 6M detector in shutterless mode (Rajendran et al., 2011). Table 20 summarizes data collection strategies for all projects in this study.

Table 20: Data collection strategies of crystallographic projects

Project	Dataset	λ(Å)	Space group	Images	Oscillation	Beam Line / Detector
PA1622	native	0.9184	$P6_1$	1500	0.1	BL14.1 / Pilatus 6M
	$HgCl_2$	1.000	$P4_12_12$	3600	0.1	X10SA / Pilatus 6M
	Thiomersal	1.000	$P4_12_12$	3600	0.1	X10SA / Pilatus 6M
	I3C-iodide	1.5418	$P6_1$	2160	0.5	BL14.2 / MX-225 CCD
	SeMet_pk	0.9760	$P6_1$	722	0.5	BL14.1 / MX-225 CCD
PA1623	native_apo		$P2_12_12_1$	100	1	BL14.1 / MX-225 CCD
	native_GSH		$P2_12_12_1$	120	1	BL14.1 / MX-225 CCD
	SeMet_pk		$P2_12_12_1$	1500	0.5	X10SA / Pilatus 6M
PA1624	native	0.9180	$P2_12_12_1$	100	1	BL14.1 / MX-225 CCD
	SeMet_pk	0.9800	$P2_12_12_1$	28800	0.1	BL14.1 / Pilatus 6M
	Yb_cocrystal	1.3853	$P2_12_12_1$	2880	0.1	BL14.1 / Pilatus 6M
PA5506	native	0.9184	$P4_12_12$	1000	0.1	BL14.1 / Pilatus 6M
	SeMet_pk	0.97953	$P6_5$	1800	0.1	BL14.1 / Pilatus 6M
	SeMet_ip	0.97992	$P2_12_12_1$	3600	0.1	BL14.1 / Pilatus 6M
	SeMet_pk	0.97981	$P2_12_12_1$	3600	0.1	BL14.1 / Pilatus 6M
	Sulfur_SAD	1.5418	$P6_5$	6800	0.5	inhouse / Saturn944+
	Zinc_hrem	1.282827	$P2_12_12_1$	3020	0.1	BL14.1 / Pilatus 6M
	Zinc_lrem	1.286153	$P2_12_12_1$	1930	0.1	BL14.1 / Pilatus 6M
PA5507	Native_1	0.91841	$P2_12_12$	240	0.5	BL14.1 / MX-225 CCD
	Native_2	0.9724	$P2_12_12$	2000	0.15	ID23-1 / Pilatus 6M

4.9.3. Data processing

All diffraction data were indexed and integrated with XDS (Kabsch, 2010) and scaled with AIMLESS (Evans & Murshudov, 2013) using the CCP4 package (Potterton et al., 2003).

XDS is an automated 3-dimensional spot integration data program running from a single input file: XDS.INP which contains all parameters required. In figure the workflow of an xds run is depicted.

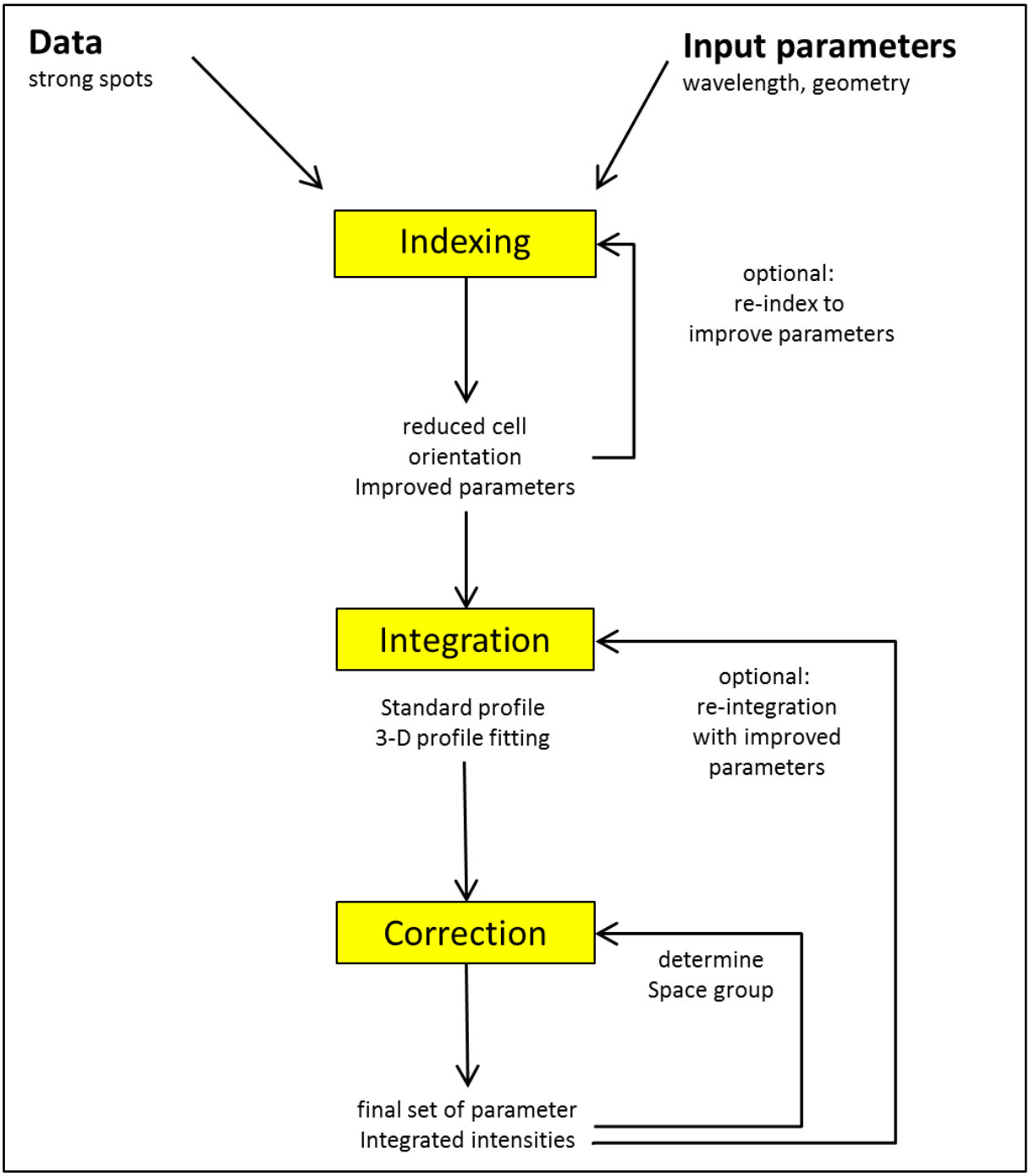


Figure 33: XDS reads measured reflection coordinates (X, Y, Z) and their indices (h, k, l) to refine the input parameter for cell parameters, beam direction, rotation axis, detector distance.

4.9.4. Structure and phase determination

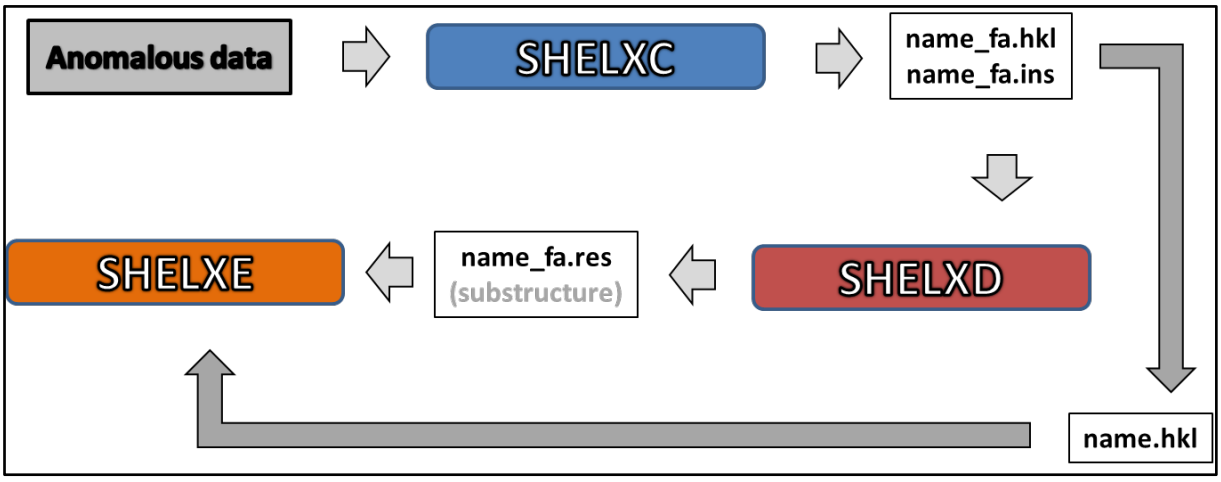


Figure 34: SHELX workflow (Sheldrick, 2010)

Phases can either be obtained by molecular replacement with e.g. BALBES (Long et al., 2008). In other cases the phases need to be determined from either multi MAD or SAD datasets collected at the selenium K-edge using SHELX (Schneider & Sheldrick, 2002; Sheldrick, 2010) to determine the phases.

Atom positions were inspected by COOT (Emsley et al., 2010). The experimental phasing pipeline AutoSol (Adams et al., 2010) was used for phasing with the located heavy atom positions (SHELXD) served as an additional input file for AutoSol (Zwart et al., 2008) implementing Xtriage for data quality analysis, PHASER (McCoy et al., 2007) for calculation of the experimental phases. RESOLVE (Terwilliger, 2000) was used for density modification and initial model-building alternated with refinement in phenix.refine (Afonine et al., 2012) with 5% of randomly chosen reflections for R_{free}.

5. Results and Discussion:

Given that two unrelated operons were selected for this study, the results are split in two; each part presents experimental results regarding one operon.

PART I

Open reading frames *PA1621-PA1624*

5.1. Structural analysis of PA1622

5.1.1. Sequence analysis of PA1622

The non-essential protein PA1622 is a 31.1 kDa gene product of unknown function. It is predicted to be localized in the cytosol of *Pseudomonas aeruginosa PAO1* (Petersen et al., 2011). Its sequence was taken from the Pseudomonas Genome Database (PGD) (Winsor et al., 2011) and analyzed by a Blast search revealing 47 orthologs in *Pseudomonas species*. Automated annotation based on the presence of conserved amino acids sequences and structural features indicated that the protein belongs to the class of hydrolases (EC 3.1.1.XX). This annotation was confirmed by a reversed position specific blast (RPSBLAST) screening for cluster of orthologous groups (COGs) as member of the alpha/beta hydrolase superfamily. Its possible function was predicted as an atropine acyl-transferase based on 35% sequence identity with the N-terminal part (residues 2-125) of an enzyme from *Pseudomonas putida,* annotated by the *Pseudomonas aeruginosa* Community Annotation Project, PseudoCAP (http://www.cmdr.ubc.ca/bobh/paap.html) (Winsor et al., 2005).

Sequence domain search predicted the residues 25 through about 125 to fold into one domain with an association to the esterase or lipase function acting on carboxylic esters (Marchler-Bauer et al., 2011). These enzymes are characterized by their catalytic apparatus involving a triad of serine as active site nucleophile with either a glutamic or an

aspartic acid residue and a conserved histidine side chain. Amongst the superfamily of alpha/beta hydrolases involving serine chemistry, the classes of p-nitrobenzyl esterases and acetylcholine esterases were reported to feature a glutamic acid side chain while other members employ an aspartic acid side chain.

Querying the PDB using the primary sequence with a sequence identity cut off set to 40% did not result any similar structure. Lowering the value to 35%, 18 structures originating from several organisms like *Pseudomonas aeruginosa* and *Rhodopseudomonas* but also *Burkholderia, Thermus spec* and *Bacillus subtilis* were identified. The results are summarized in table 1. Lowering the cut off to 25% revealed 112 unique entries. The protein sequences were extracted and analyzed by multiple sequence alignment (Sievers et al., 2011). PA1622 clusters with a thioesterase domain involved curacin A biosynthesis (Gehret et al., 2011) as well as 2 proteins, both reported with a proline iminopeptidases activity, from *Xanthomonas* (Medrano et al., 1997) and *Thermus species* (Okai et al., 2007), highlighted in bold and marked with asterisk in table 21.

Table 21: Results from BLAST search against the Protein Data Bank (Altschul et al., 2005).

Protein	Origin	Sequence identity (%)	Query coverage (%)	E-value	Pdb code
Thioesterase Domain *	***Lyngbya majuscula***	**28**	**86**	**3E-23**	**3QIT**
Alpha/Beta hydrolase	*Nostoc sp PCC7120*	35	34	2E-09	3QYJ
Epoxide hydrolase	*Pseudomonas aeruginosa PAO1*	37	34	8E-09	4BAU
Fluoroacetate dehalogenase	*Rhodopseudomonas palustris*	35	37	3E-08	3R3V
Acyltransferase	*Salmonella enterica*	33	34	6E-08	4NVR
E-2-(Acetamidomethylene) succinate hydrolase	*Mesorhizobium loti*	31	57	1E-07	3KXP
Aryl esterase	*Burkholderia cenocepacia*	32	40	2E-07	4X00
Chloroperoxidase L	*Streptomyces lividans*	33	42	9E-07	1A88
Proline iminopeptidase *	***Xanthomonas campestris***	**32**	**35**	**2E-06**	**1AZW**
Epoxide hydrolase B	*Mycobacterium tuberculosis*	32	38	2E-06	2E3J
Bromoperoxidase A2	*Streptomyces aureofaciens*	26	63	4E-06	1BRO
Hsad	*Mycobacterium tuberculosis*	32	41	5E-06	2VF2
Chloroperoxidase T	*Streptomyces aureofaciens*	30	39	1E-05	1A7U
Proline iminopeptidase *	***Thermus thermophilus***	**35**	**32**	**2E-05**	**2YYS**

5.1.2. Structural prediction

Structural prediction was carried out with Phyre (Kelley & Sternberg, 2009). The resulting model is covering 273 residues (95%) of PA1622's total sequence. A liver epoxide hydrolase from *Mus musculus* (PDB: 1CR6) (Argiriadi et al., 1999) served as template. The confidence score of 100% indicates an accurate model (figure 35A and B).

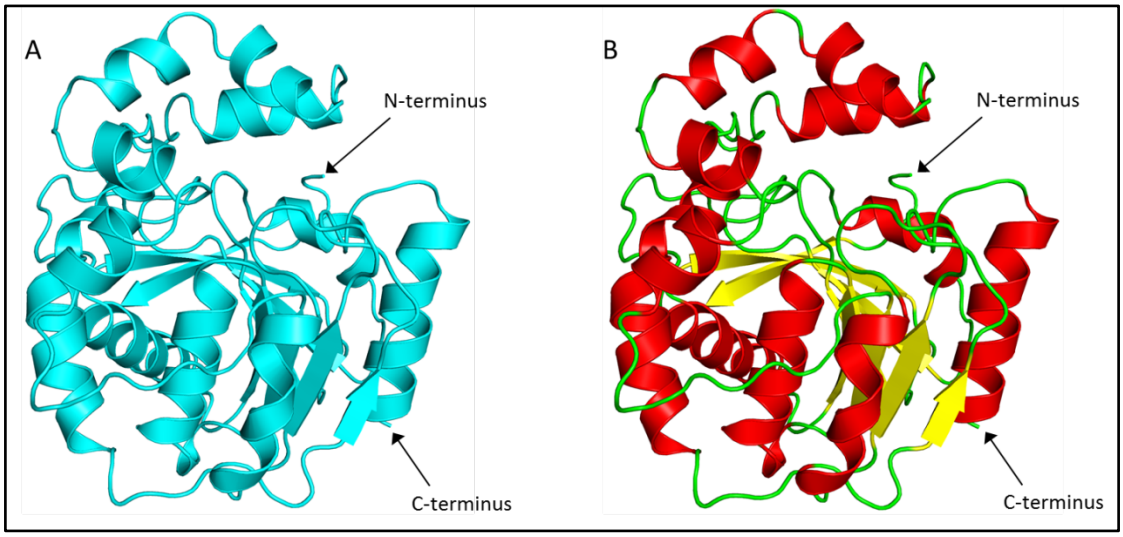


Figure 35: Model generated by Phyre structural prediction (Kelley & Sternberg, 2009).

Besides this, 19 similar models have been generated based on different PDB templates (table 21). All 20 structures are shown in its superposition in figure 36.

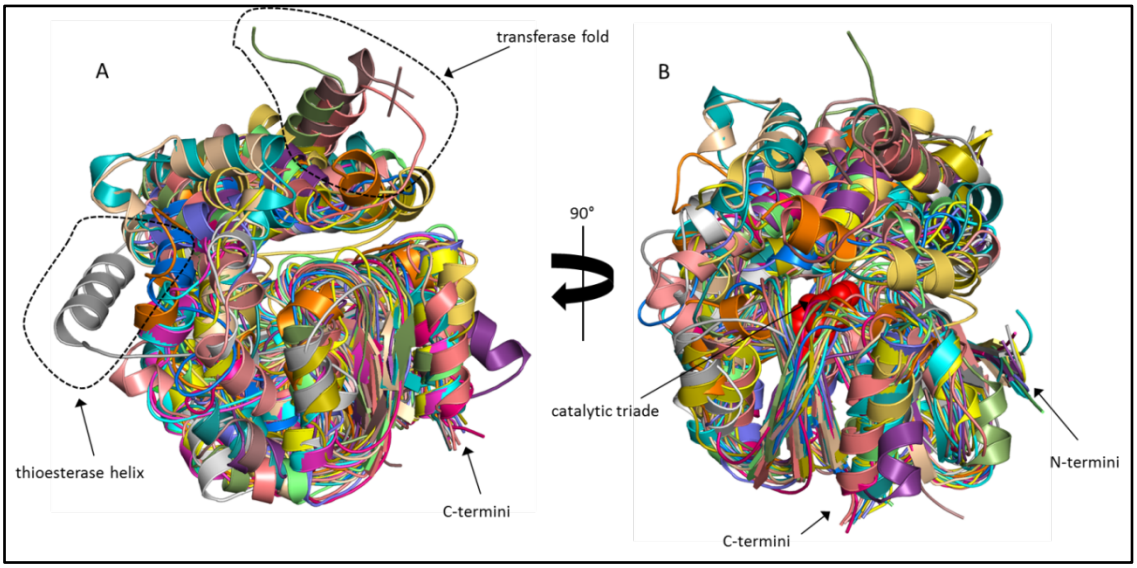


Figure 36: Superposition of all protein models generated by Phyre prediction. All chains are colored different. Observed differences are highlighted with dashed circles. The N- as well as the C-termini are clustered and marked by arrows.

Table 22: Summary of protein models that were generated by Phyre. Results are ranked according to the TM-score reported by Phyre (Kelley & Sternberg, 2009). Epoxide hydrolases are marked in bold and transferases in italic.

TM-score *	Ranking	PDB id (figure 3)	(potential) function	PDB header
1.00	1	1cr6 (Chain 'A')	soluble epoxide hydrolase	Hydrolase
0.97	2	3i28 (Chain 'B')	soluble epoxide hydrolase	Hydrolase
0.78	13	3kda (Chain 'C')	epoxide hydrolase	Hydrolase
0.77	19	4inz (Chain 'D')	epoxide hydrolase	Hydrolase
0.75	18	4b9a (Chain 'E')	putative epoxide hydrolase	Hydrolase
0.74	17	3wib (Chain 'F')	haloalkane dehalogenase	Hydrolase
0.73	15	2e3j (Chain 'G')	epoxide hydrolase (ephb)	Hydrolase
0.72	20	1c4x (Chain 'H')	Carbon-carbon bond hydrolase	Hydrolase
0.72	12	4opm (Chain 'I')	putative lipase (lip1)	Hydrolase
0.71	16	1b6g (Chain 'J')	haloalkane dehalogenase	Dehalogenase
0.70	8	3oos (Chain 'K')	alpha/beta hydrolase	Hydrolase
0.69	10	2r11 (Chain 'L')	putative hydrolase (2632844)	Hydrolase
0.68	9	* 3qit (Chain 'M')	thioesterase domain	Hydrolase
0.67	5	4d9j (Chain 'N')	designed tetrahedral protein cage	de novo protein
0.65	11	4ose (Chain 'O')	putative hydrolase	Hydrolase
0.63	14	2b61 (Chain 'P')	O-acetyltransferase	Transferase
0.62	3	1azw (Chain 'Q')	proline iminopeptidase-like	hydrolase
0.62	7	2vav (Chain 'R')	acetyltransferase	Transferase
0.61	6	d1wm (Chain 'S')	Proline iminopeptidase-like	hydrolase
0.61	4	4qlo (Chain 'T')	homoserine o-acetyltransferase	Transferase

The results are ranked according to their TM-score, which represents the structural overall similarity. Interestingly, six out of the seven best models are based on proteins with an epoxide hydrolase function, followed by hydrolases with lipase-, haloalkane- or unknown function. Proteins with transferase and peptidase activity complete the table. Although the structures share the same overall fold, differences can be identified and

some of the structures categorized in protein sub-folds (figure 37). The thioesterase domain (asterisk in table 22) shows a different fold (colored grey - figure 36A and 37C) as well as the transferase based models (rank 4, 7 and 14, table 22) and the proline iminopeptidase-like have a partially different secondary structure element organization (figure 37A and 37B).

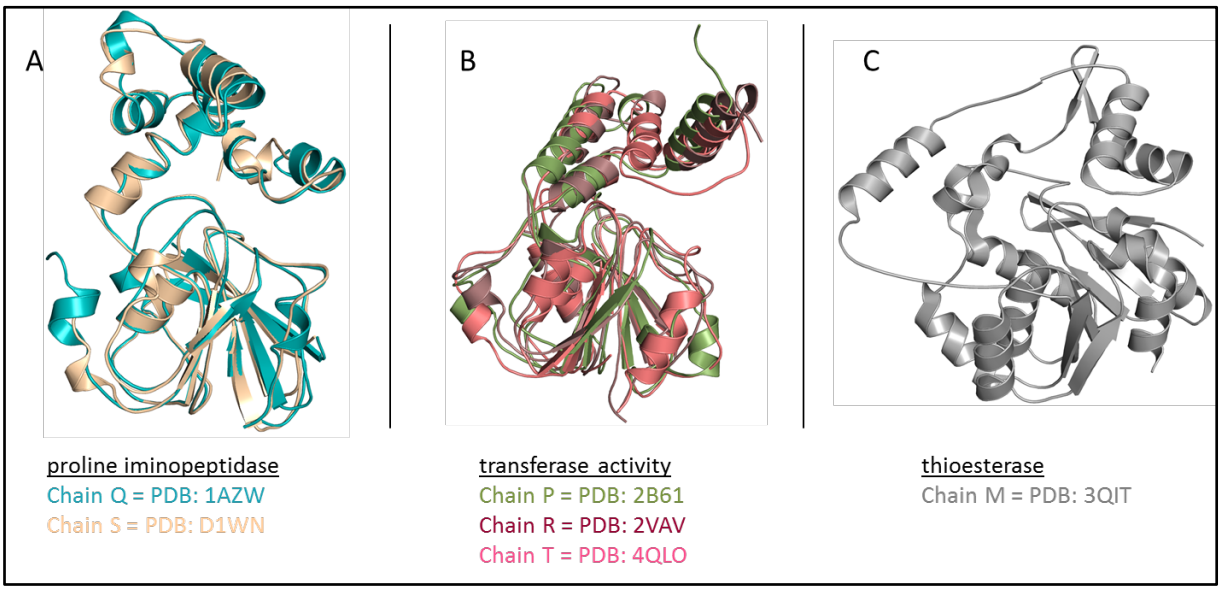


Figure 37: Structural 'outliners' amongst the Phyre results. Chains are colored and the pdb code is given according to each chain.

5.1.3. Structure of PA1622

Selenomethionine labeled as well as native PA1622 was crystallized at 30 mg/ml in space group $P6_1$ with almost identical cell dimensions (table 2). Final buffer conditions are summarized in table 3 and 4. The content of the asymmetric unit was calculated for four to six monomers with a Matthews coefficient of 3.16, 2.52 and 2.1 reflecting a solvent content of 61%, 51% or 41%, respectively. The 286 amino acid protein features 7 methionine residues within a single protein chain. The structure was solved using SeMet SAD phasing followed by phase extension with native data to higher resolution. SHELX (Sheldrick, 2010) located the heavy atom sites and the phenix application AutoSol (Zwart et al., 2008) was used to solve the structure including initial model building. Visual inspection indicated the presence of four protein molecules in the asymmetric unit assembled to a homo-tetrameric complex, a dimer of dimers. The crystal symmetry is shown in figure 38. The monomer of PA1622 is colored in yellow and its 6-fold symmetry

equivalent molecules are highlighted in red; the next corresponding monomer is colored in light yellow.

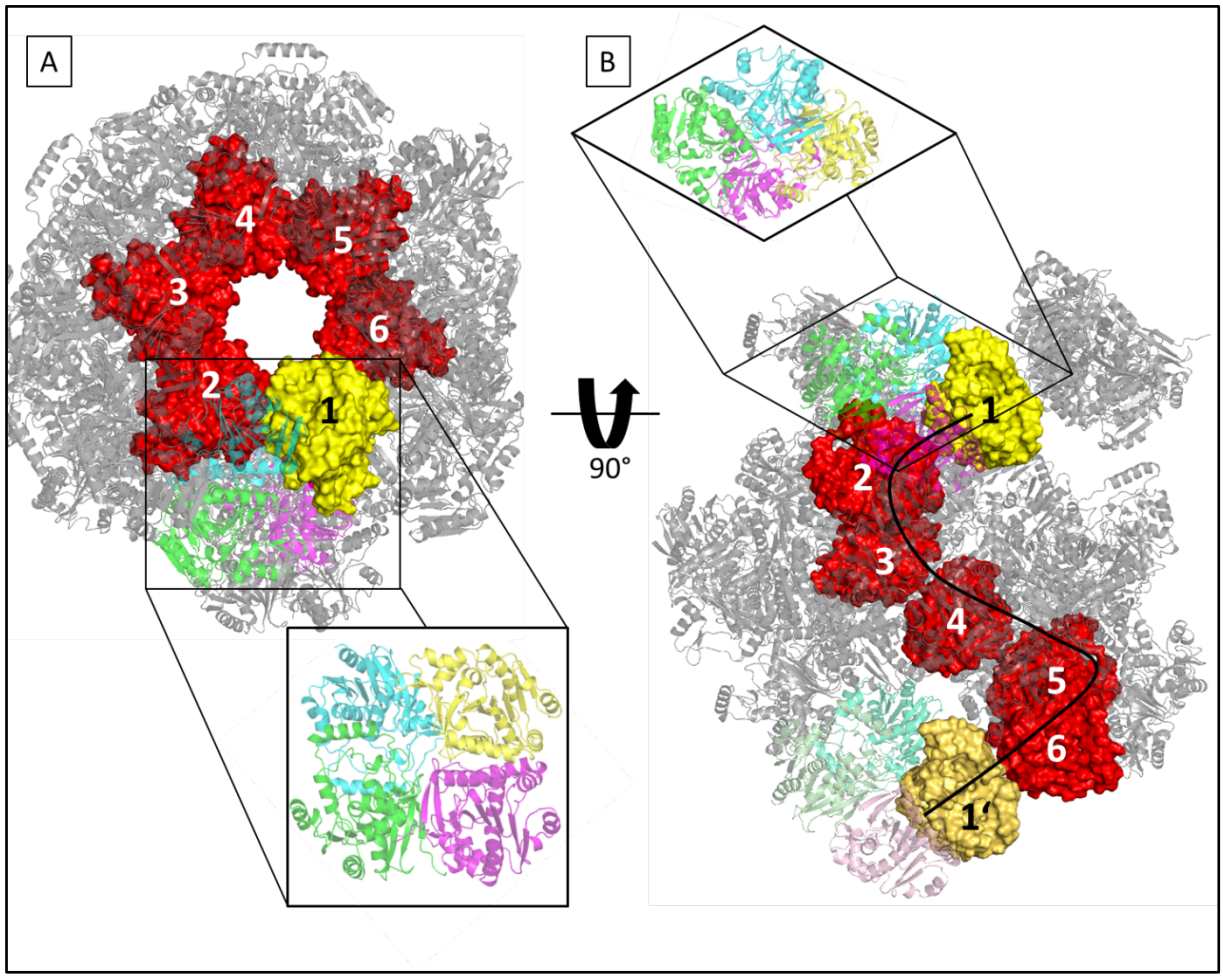


Figure 38: Crystal organization of PA1622 native crystals in space group $P6_1$ with a tetramer arrangement of PA1622. A top shows the six fold symmetry (A) and a side view (B). The tetramer is always shown in identical orientation aside.

Table 23: Crystallization conditions of native PA1622, crystals grew within 3 days to a size of about 450 µm.

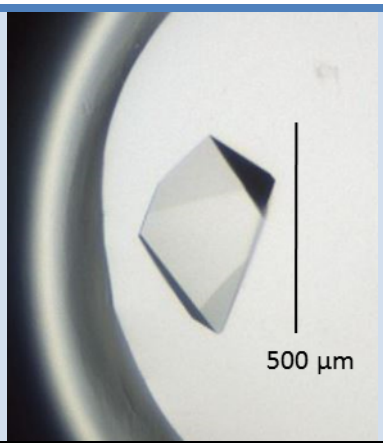
State	(initial) Condition	Picture	Time
Native	(Midas B11) 20% PEG 400 10% 1-Propanol		3 days

Table 24: Pure selenomethionine labeled protein crystallized within 7 days to a size of 200 µm in the following condition.

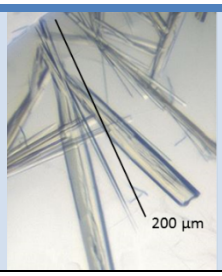
State	(initial) Condition	Picture	Time
Selenomethionine labeled	(JCSG+) 20% PEG 3350 0.2 M Mg-formate		7 days

Table 25: Data collection stastics for PA1622

Data collection	PA1622 SeMet [α #]	PA1622 native [α #]
Detector	Rayonix MarCCD	Pilatus 6M
Wavelength (Å)	0.9760	0.91841
Resolution range (Å)	20 – 2.8 (2.92 – 2.8)	45.5 – 2.1 (2.14 – 2.1)
Space group	$P6_1$	$P6_1$
Unit cell parameters (Å)	136.3 136.3 148.2	136.8 136.8 148.1
Unit cell angles (α, β, γ)	90° 90° 120°	90° 90° 120°
Total No. of measured reflections	895996 (109324)	790175 (39646)
Unique reflections	38348 (4664)	91396 (4503)
Multiplicity	23.4 (23.4)	8.6 (8.8)
ano. Multiplicity	11.8 (11.8)	--
Mean I/σ(I)	14.8 (4.2)	14.8 (2.2)
Completeness (%)	99.7 (99.9)	100 (100)
ano. Completeness (%)	99.7 (99.9)	--
Average Mosaicity [°]	0.11	0.06
R_{merge} (%)	29 (97.8)	9.7 (108)
R_{meas} (%)	29.6 (100)	10.3 (114)
R_{pim} (%)	6.1 (20.5)	3.5 (38.6)
CC(1/2) [°]	0992 (0.894)	0.999 (0.706)

[#]Data were collected from single crystal; [+]Values in parentheses refer to the highest resolution shell.
[α]Data collected at at BESSY II, BL14.1 (Mueller et al., 2012), [°]Mosaicity and CC(1/2) reported by XDS (Kabsch, 2010; Karplus & Diederichs, 2012), [$]Rpim = Σhkl (1/(N – 1))1/2 Σi | Ii (hkl) – < I (hkl) > | / Σhkl Σi Ii (hkl), (Weiss, 2001).

The polypeptide chain of each monomer was visible and ordered to the C-terminus. Manual adjustments and improvements of the initial model were done using COOT (Emsley & Cowtan, 2004) alternating automated refinement and phase improvement

carried out by the phenix application refine (Afonine et al., 2012). The model was refined to a R_{work} of 17.0% and R_{free} of 19.7%. The refinement statistics are summarized in table 26. For all amino acids except the first two residues electron density could be observed and interpreted by the polypeptide chain.

Table 26: Refinement statistics of the native dataset of PA1622

Data collection	PA1622 SeMet $^{\Phi}$	PA1622 native
Resolution range (Å)	20 – 2.8 (2.92 – 2.8)	45.5 – 2.1 (2.14 – 2.1)
Unique reflections	38322 (3826)	790175 (39646)
R_{work} (%)	25 (33.6)	17 (23.7)
R_{free} (%)	28.7 (37.9)	19.7 (25.7)
No. atoms	8326	9030
Protein *	8326	4507
Water	--	475
Ligands	--	48
Protein residues	1104	1134
Average B-factors		
Protein	17.7	46
Water	--	41
R.m.s. deviations		
Bond length (Å)	0.004	0.003
Bond angle (°)	0.94	0.71
Rotamer outlier (%)	0.88	0.24
Ramachandran plot (%)		
Favored region	92	96.0
Outlier	1.6	0.27
Validation score		
Clashscore	9.1	0.47
MolProbity$^{\&}$	1.97 (99th)	0.94 (100th)

*** indicates the number of non-hydrogen, non-solvent atoms; $^{\&}$As reported by MolProbity (Chen et al., 2010), http://molprobity.biochem.duke.edu/); ϕ the selenomethionine data set was left non refined.**

The overall structure is shown in figure 39. The four chains arrange in a flat donut-like structure leaving a whole in the middle (figure 39B & C).

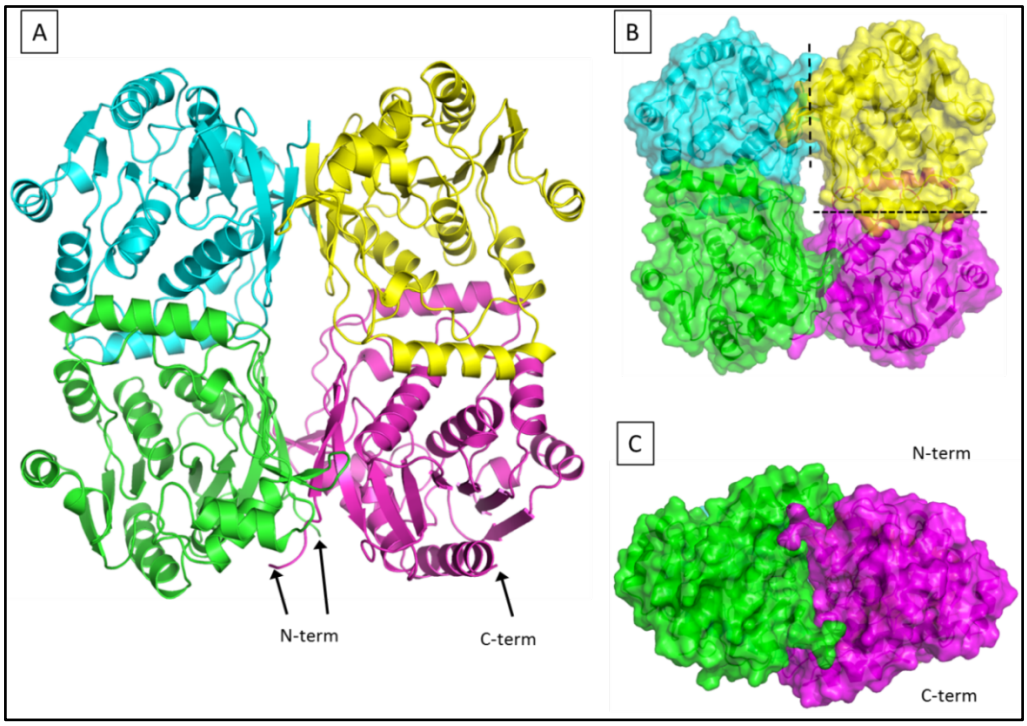


Figure 39: Overall structure of PA1622 with differently colored chains (A). The surface representation shows a hole in the middle of the tetrameric complex (B). Dimer interfaces are indicated by dotted lines in B.

The protein PA1622 possesses, as expected, α/β-hydrolase fold with ten beta sheets and nine alpha helical main secondary structural elements (figure 40).

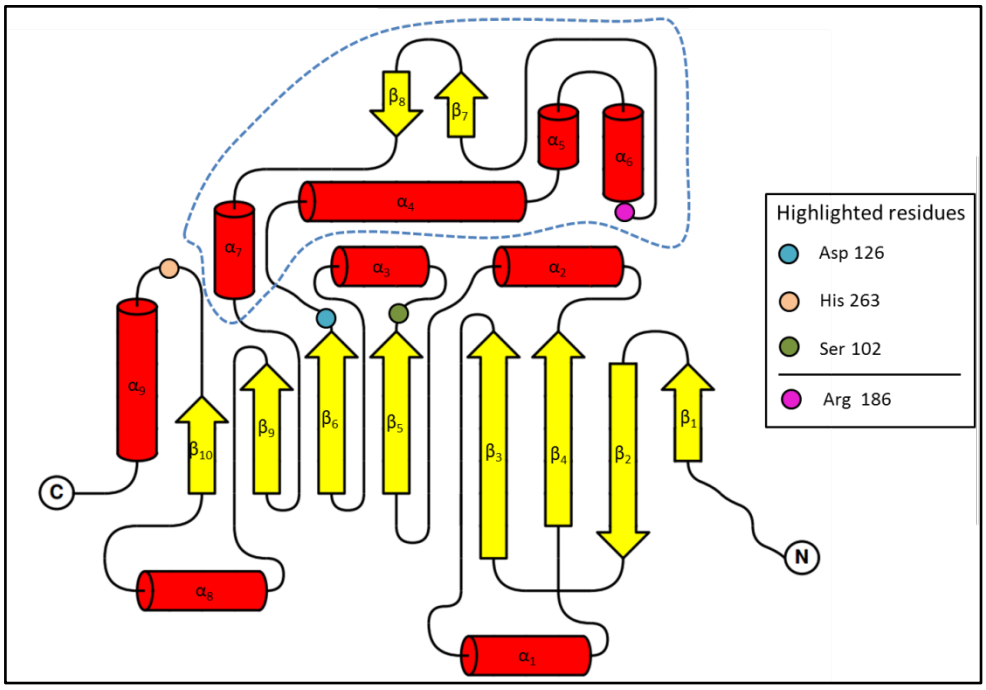


Figure 40: Overall topology of the secondary structure elements of PA1622. A dashed line surrounds the smaller domain.

The structure of a monomer can be clearly separated in two domains of different size and overall morphology. The core (residues 1-127 and 215-286 (colored in light blue, figure 41B) is built by an arrangement of 8 mostly parallel beta sheets while β_2 is the only one running antiparallel. The blade of sheets is tightly twisted by about 120° from the first (β_1) to the last one (β_{10}) finishing the organization close to the C-terminus. It is flanked by alpha helices α_1 and α_9 on one side while alpha helix 2, 3 and 8 packs on the other side of the large beta sheet. A smaller domain is inserted between amino acid 128 and 214. It comprises four alpha helices and one small anti parallel beta sheet (figure 41B, colored in light yellow) sitting on top of the compact core-domain like a covering lid collected via linker regions stained in green (figure 41A-C).

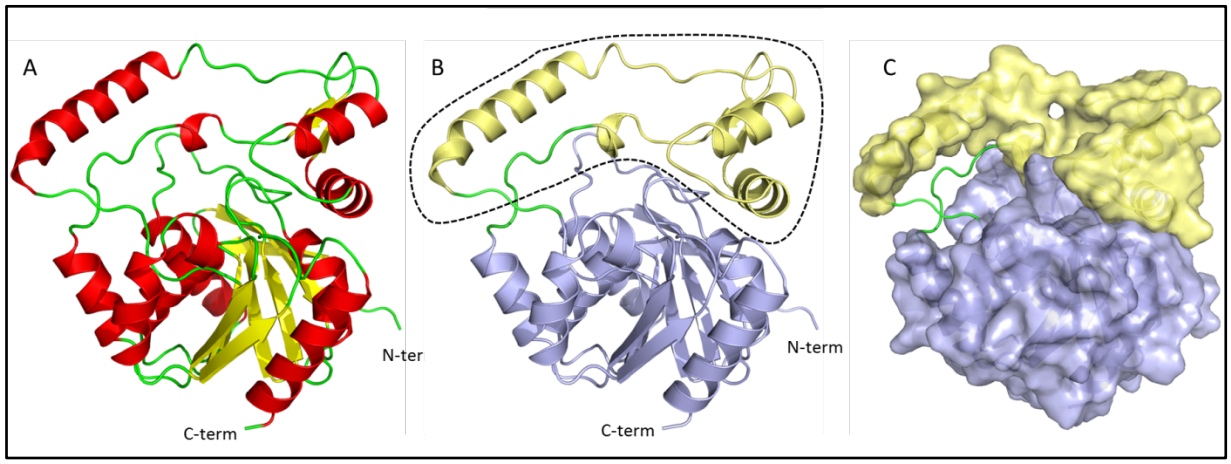


Figure 41: Structure of the PA1622 monomer showing the two domains colored in yellow and light blue (B). The secondary structure elements are shown in A with beta sheets colored in yellow and in red the alpha helices. A surface representation of both domains is shown in C.

The rather compact structure features one helix, alpha helix 4 (residues 135-152), which is the 'protruding helix'. It is the only structural element that doesn't have direct contact to the core domain and only a few interactions with the lid domain itself.

5.1.4. Domain organization

Superposition of the four protein chains demonstrated an identical position of the lid in respect to the core domain (root mean square deviation = 0.24 Å for 286 C_α). The position of the lid is maintained by mostly hydrophobic contacts of alpha helix 6 with several loop regions of the core domain. An arginine residue (R186) located at on helix 6 (figure 40) anchors the lid- to the core domain via a specific salt bridge to a glutamate side chain (E7)

located in the N-terminus of beta sheet 1. Besides this an aspartate side chain (Asp59) is creating a hydrogen bond to the backbone amide of alanine 40 and asparagine 39 makes contact to the backbone carbonyl of leucine 185, both contributing to the stability. Additionally, the lid domain is participating in dimer interaction.

This protein fold leads to 88% of all protein residues exposed to the surface with a calculated total area of about 13800 A² per monomer (Krissinel & Henrick, 2007). Domain analysis indicates that a dimer of two chains and the tetrameric assembly are possible assemblies that are favored over the monomeric state (table 27).

Table 27: Analysis results of the PA1622 protein tetrameric complex using PISA (Krissinel & Henrick, 2007)

Composition	stable	Surface area (sq. Å)	Buried area (sq. Å)	ΔG^{int} (kcal/mol)	ΔG^{diss} (kcal/mol)
ABCD	yes	42290	11230	-78.4	0.4
AB	yes	23540	3320	-35.7	22.8
CD	yes	23320	3340	-34.3	21.6
A	Yes	13836	--	--	--

The solvent accessible surface is reduced by about 15% upon protein dimerization and further reduced by assembly of the tetramer with 27% of the total accessible surface buried. The gain in solvation free energy was calculated to -35 kcal/mol and -78 kcal/mol, respectivey. The dimer interface between chain A - B or C – D, generated by alpha helix 2-4 and 7 as well as the loop regions connecting β_4 and α_2 and β_8 with α_7, is mostly hydrophobic with a calculated interface area of 1670 Å². The second interface between the two dimers (figure 42C) is rather hydrophilic with a calculated size of 1330 Å². Generated by alpha helix 2, 5 and 6 it also involves beta sheet 1, 2 and the beta-turn ($\beta_7\beta_8$) patching to the other monomer's beta blade, extending it to a 10 stranded beta sheet. Furthermore the N-terminus grabs the other monomer in a finger like manner (figure 8C). The dominating interactions here are eight salt bridge- and five hydrogen bonds across the interface. The P-value for the calculated solvation free energy gain (-5.3 kcal/mol) with 0.9 is rather large, indicating that this interface might be an artifact of crystal packing (Krissinel & Henrick, 2007; Krissinel, 2010). This is supported by a low CSS value of 0.2. Both interfaces do contribute to the proteins stability as calculated by PISA (Krissinel, 2011). To enlighten this, DLS data were collected at different temperatures ranging from 20°C to 5°C to exclude temperature artifacts (table 28).

Table 28: Results of DLS experiment using PA1622; The translational diffusion coefficient = *D*, the hydrodynamic radius (*R*) with the corresponding diameter, *Dia*. The polydispersity (Pd) is calculated to a percentage value. Molar mass (*MW-R*) is estimated upon particle conformation, size and solution density. All calculation were done based on a classical spherical fitting model (Nobbmann et al., 2007).

Temperature (°C)	Peak (#)	D (cm^2/s)	R (nm)	Dia (nm)	Pd (%)	MW-R (kDa)	Int (%)	Mass (%)
20	1	5.76E-07	3.6	7.3	9.4	**69.0**	98.81	*99.9994*
	2	1.72E-08	122.1	244.3	15.5	**2571.0**	1.19	0.0006
15	1	5.72E-07	3.7	7.3	10.4	**70.7**	100	*100*
10	1	4.40E-07	3.5	7.0	5.9	**63.7**	100	*100*

Data revealed a molecular particle weight of approximately 65 kDa with a diameter of 73 Å. The diameter of a protein in the crystal structure dimer (62 kDa) and tetramer (248 kDa) were measured to approximately 75 Å and 90 Å, respectively. The low degree of polydispersity and a mass accuracy of 100% indicate a homogenous monodisperse solution. In summary, the tetramer is most likely a crystallization artifact rather than a natural stable structure (table 8), which is also in congruence with results from gel filtration where PA1622 eluted as a dimer.

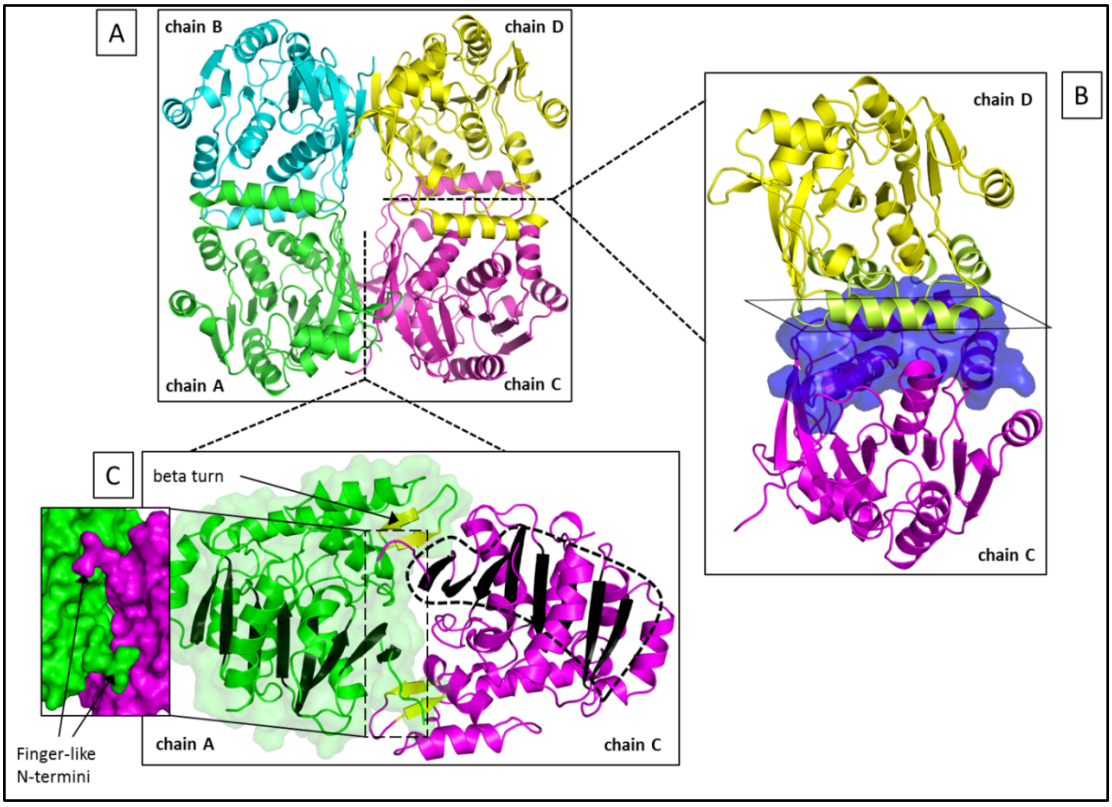


Figure 42: Dimerization interfaces of PA1622. Chain C and chain D (magenta and yellow) share the hydrophobic interface while chain C and chain A have a polar interface. The 8 fold beta blade is colored in black with the extended by the 2 fold beta turn (yellow) of the other monomer.

Dimer protein interface was used as subject searching for similar interfaces within the *PDB* archive (Berman et al., 2000), one similar dimer assembly is revealed, a thioesterase domain from curacin biosynthetic pathway, CurM (PDB: 3QIT, (Gehret et al., 2011)) already identified by sequence alignment (table 21). The asymmetric unit of 3QIT contains four protein chains as well but organized in two separate dimers instead of the tetrameric structure.

5.1.5. Structural comparison to similar predicted structures

Superposition of PA1622 with all predicted models by Phyre (figure 9A) revealed one structure of significant similarity. The model previously named as 'outlier protein fold' since it did not match the other 19 predicted structures resembles the *Pseudomonas* protein PA1622 while the model as predicted best by Phyre (figure 43A and B) does not.

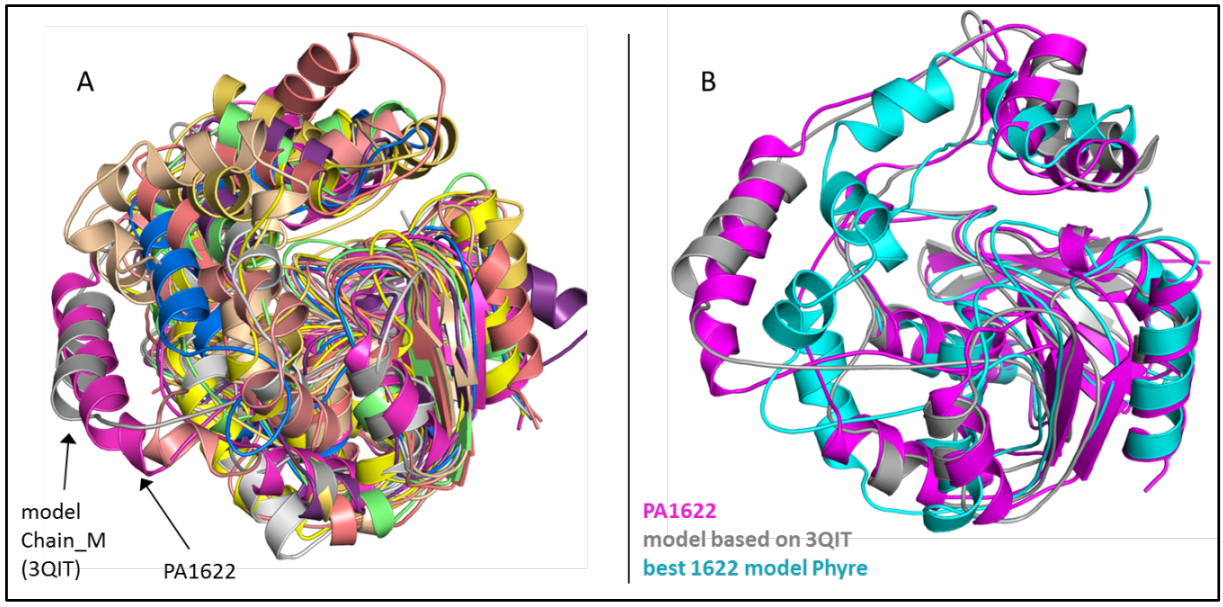


Figure 43: Superposition of the structure of PA1622 to all Phyre models (A) reveals the model M (3qit) as best (B).

The thioesterase domain monomer superimposes with an r.m.s.d. of 1.9 Å and the protein dimer with 2.3 Å to PA1622 (figure 44A and B). The hydrophobic protein dimerization interface has significant similarity to the one in PA1622. Further significant similarities to other dimer interface as well as other predicted model could not be detected.

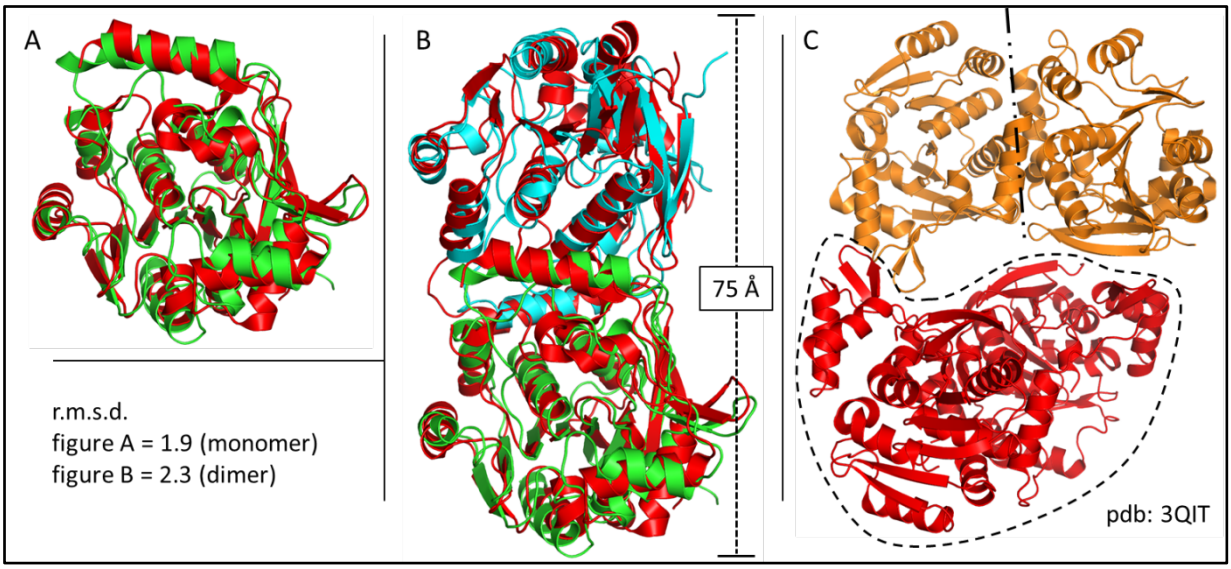


Figure 44: Superposition of PA1622 and 3QIT. A, the monomers of both superpose with an root mean square deviation (rmsd) of 1.9 Å and dimers (B) align with 2.3 Å rmsd. The protein dimer is marked by the dashed circle. The dashed straight line indicates the hydrophobic dimerization interface similar to the one in PA1622 (fig. 8B).

5.1.6. Active site of PA1622

The putative active site of PA1622 was identified through sequence homology with other proteins of similar fold. An alignment of 1000 non-redundant protein sequences of members of the alpha/beta hydrolase superfamily revealed a conservation of the catalytic histidine with either a serine or an aspartic acid as active site nucleophile. The position of the third amino acid of the catalytic triad shows a larger number of possible residues while aspartic acid is the most abundant amino acid. The more conserved core-domain comprising the key amino acids can be clearly distinguished from the lid domain, which is rather composed of random amino acids (figure 45).

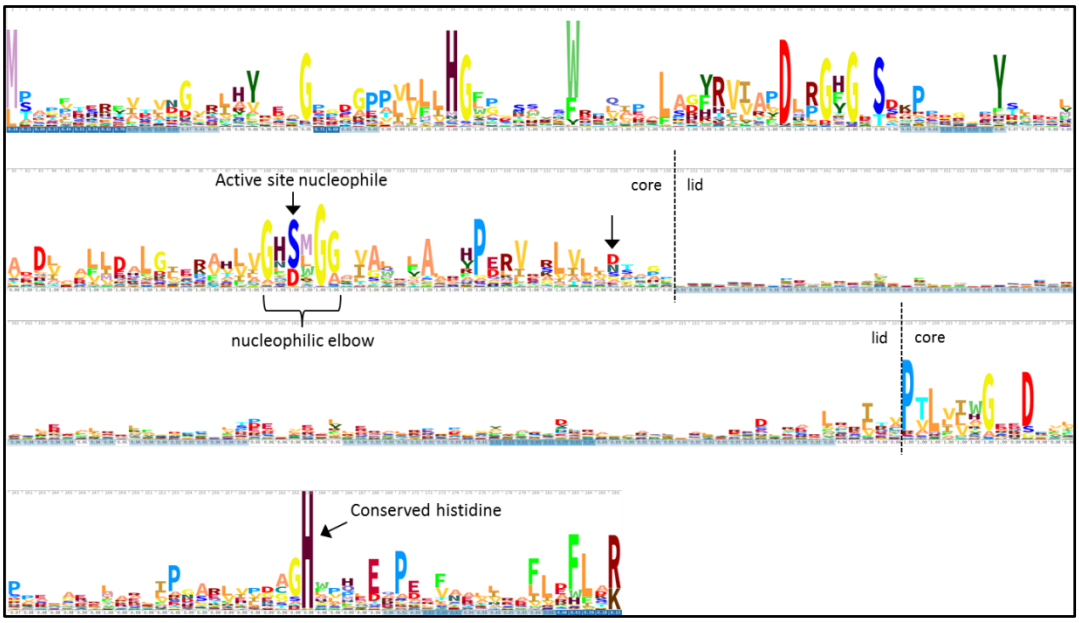


Figure 45: Sequence alignment of 1000 non-redundant protein sequences of recognized members of the alpha/beta hydrolase superfamily. The figure was generated by Skylign (Wheeler et al., 2014).

The active site of PA1622 contains a catalytic triad of Ser-His-Asp and is located in a cleft between the lid- and the top of the core domain. The nucleophilic serine residue is found within a signature sequence Gly_{100}-His_{101}-**Ser**$_{102}$-X_{103}(M)-Gly_{104} (figure 45) in the constrained geometric nucleophilic elbow which is a hallmark of alpha/beta hydrolase proteins. It is located at the N-terminus of an alpha helix (α_3) which supports the stability of the deprotonated state during catalysis by helical dipole stabilization. The histidine residue (His263) and the aspartic acid residues (Asp126) are both located on loop regions connecting β_{10} to α_9 and β_6 to the protruding helix, α_4. Particularly, the histidine of the signature sequence (His101) makes a hydrogen bond to the backbone carbonyl of the active site histidine, stabilizing the alignment of the triad in close proximity with a distance of 2.7 Å from serine to histidine and 2.6 Å from histidine to the aspartic acid (figure 46).

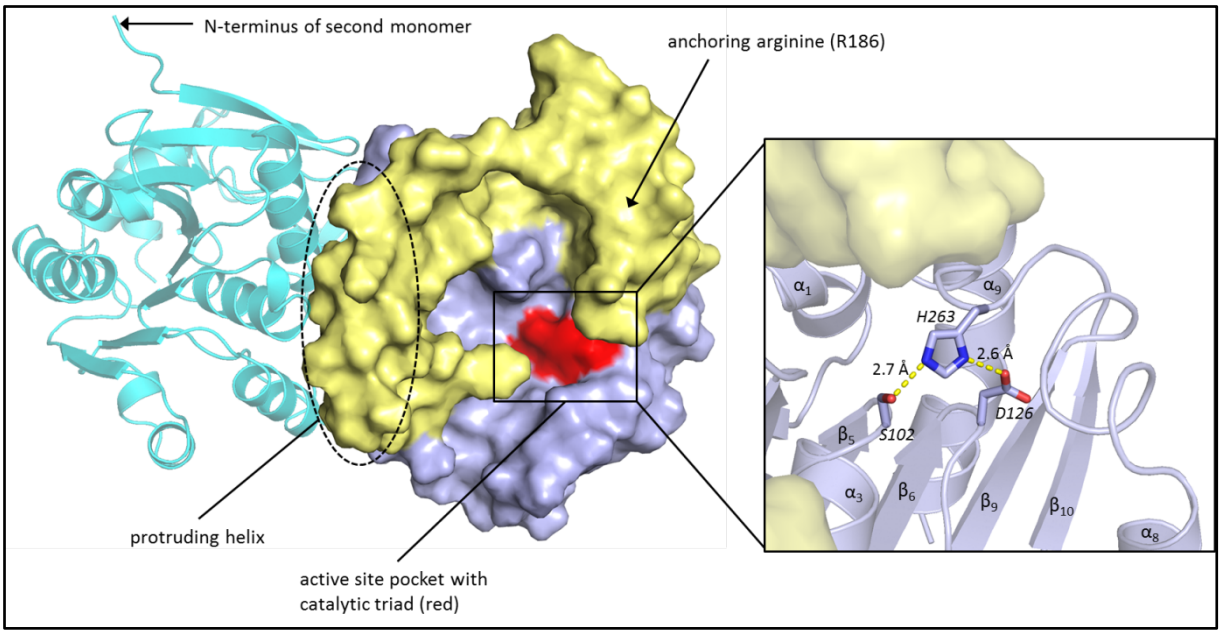


Figure 46: Active site pocket of PA1622. The catalytic triad colored in red in the overall picture is shown in the close up in greater detail (Ser 102, Asp 126 and His 263).

The largest part of the active site pocket in PA1622 is lined with aliphatic residues including Ile106, Trp36, Leu77, Trp78, Met103, Pro130, Leu212, rendering the largest portion hydrophobic except for 3 polar residues lining the other side of Ser102. An aspartate (Asp38) and two histidine residues (His101 and His264) are located in closest proximity to the catalytic triad. The amide backbone atoms of methionine (Met101) and tryptophan (Trp36) generate an oxi-anion hole to compensate distributed charges of an intermediate state during catalysis. In fact, while refining the protein structure, extra electron density which could not be interpreted by the polypeptide chain could be observed. This density is rising from the active site, covalently modifying the serine residue and branches out in the active site pocket. However, the molecule could not be full interpreted due inconsistencies in electron density. The ligand molecule possibly contains a polar substituted ring structure, which perfectly serves for pi-stacking interaction with tryptophan (Trp36) already flipped into position (figure 47) sitting 3.2 Å above the ring. Substituted ring charge is coordinated by the backbone carbonyl-atoms of three neighboring residues leucine 205, threonine 206 and leucine 207. The contribution of the carbonyl backbone of leucine 205 is mediated by a coordinated water molecule. The other end of the ligand is likely rather polar and located in the hydrophilic part of the active site pocket. It is coordinated by hydrogen bonds to the side chains of histidine 264,

aspartate 38 as well as the carbonyl backbone of tryptophan 36. Extra density was also observed in the oxi-anion hole, indicating that a possible intermediate step was trapped in the crystal. The fact that none of the compounds present in the crystallization condition explains this density leads to the conclusion that the ligand must have been co-purified. However, mass spectrometry analysis using dissolved crystals or purified protein did not show any change in mass, compared to the calculated value. This is probably due to the loss of the modification during the ionization process of the protein sample. A possible structure of the ligand is indicated by a white line in figure 47.

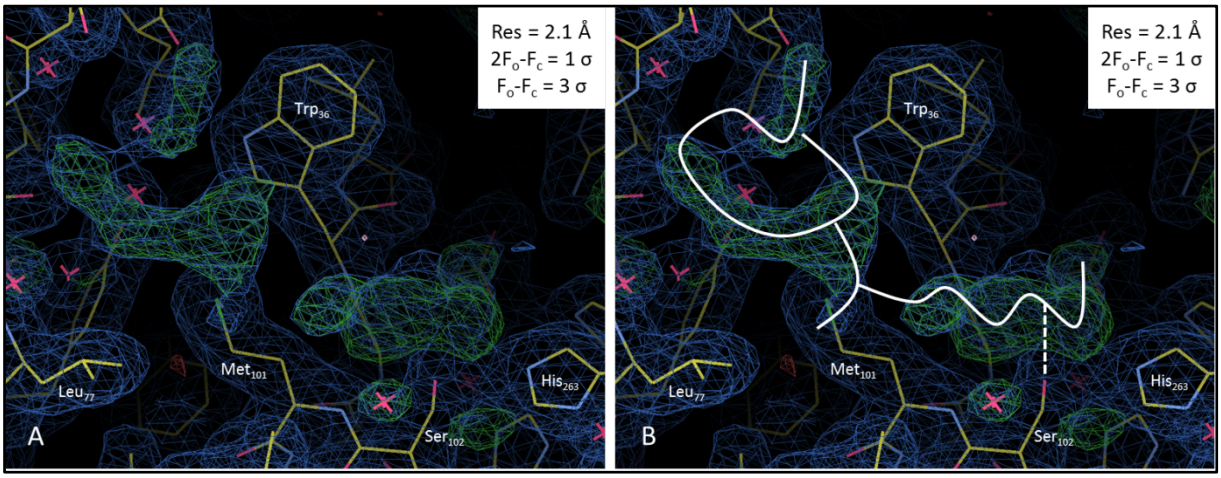


Figure 47: Extra electron density was observed in the active site pocket of PA1622 (A). This density corresponds to a larger ligand moiety (B). The dashed line indicates the covalent bond to the active site nucleophile.

5.1.7. Homologs regulate activity involving lid domain movement

Notably, the active site of PA1622 is widely exposed to the surrounding environment and not shielded by the lid domain as reported for other similar proteins (Biswal et al., 2008; Lack et al., 2008; Claxton et al., 2009). The lid domain was shown to be a regulatory unit of the protein's activity in a structural homolog, adopting two distinct conformations (Sun et al., 2014). In the open, inactive conformation the residues assembling the catalytic triad are distant from each other with a widely opened active site pocket. A global conformational change involving a reorientation of the lid domain causes a drastic shrinkage of the active site pocket, bringing the catalytic into close proximity.

This however, does not apply to PA1622, because its dimeric organization PA1622 fixes the lid-domain in an open position that does not allow a lid movement in the same way as described for a similar *Pseudomonas aeruginosa* protein with lipase activity (Nardini et

al., 2000; Cherukuvada et al., 2005) or the *E.coli* protein MenH (Sun et al., 2014). Although the core domain structures are rather similar, the structural composition of the lid domain as well as its conformation shows major differences.

5.1.8. Structure in context

Search for similar structures was carried out by PDBeFold (Krissinel & Henrick, 2007) and resulted in five hits with significant structural similarity (table 29).

Table 29: Search result from PDBeFold sorted by Q-score: Five structures have been identified.

##	Q-score	P-score	Z-score	RMSD (Å)	N_{algn}	Nucleophile	Seq-id (%)	PDB ID
1	0,55	14,46	11,88	2,12	249	Ser106	0,27	**3QIT**
2	0,54	19,29	13,87	1,71	233	Ser106	0,23	3KXP
3	0,41	8,421	10,12	2,38	233	Ser104	0,22	3BWX
4	0,39	10,45	10,6	2,35	225	Ser154	0,2	3P2M
5	0,38	8,154	9,94	2,22	221	Ser156	0,28	3C5W

The two structures ranked 4 and 5 can be grouped into one sub-class with slightly different fold (Xing et al., 2008; Zheng et al., 2011). Structures ranked second and third possess an N-terminal extension and an altered fold compared to PA1622 (McCulloch et al., 2010). The result ranked best is the previously discussed protein domain with thioesterase function and the only one that matches all secondary structure elements (figure 48B) (Gehret et al., 2011).

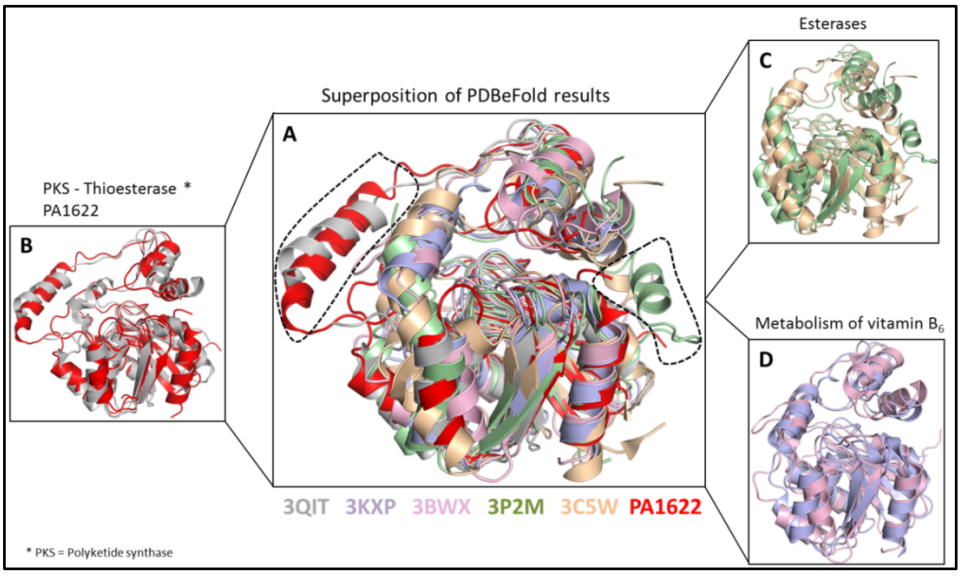


Figure 48: Superposition of PDBeFold results revealing one structure as closest structural homolog.

Results from a DALI search (Holm & Rosenström, 2010) (table 30) analyzed by multiple sequence and structure alignment revealed an insertion of 22 amino acids into the primary sequence of PA1622 and 3QIT (figure 49).

Table 30: Protein structures identified as structural homologous through a DALI search. The best-ranked 15 are listed.

PDB ID	Z-Score	RMSD (Å)	Seq Id (%)	Protein function (as known)
3qit	29,2	2,4	27	Polyketide synthase thioesterase
3kxp	29,1	2,5	22	α-(N-Acetylaminomethylene)succinate hydrolase
4ose	26,6	2,5	18	Putative hydrolase
3bwx	25,4	3	22	Alpha/beta hydrolase
3c5v	25,3	2,9	26	Phosphatase methyltransferase 1
1j1i	25,2	2,6	17	Meta cleavage compound hydrolase
3qyj	24,7	3	22	Putative Alpha/beta hydrolase
1y37	24,5	3	22	Fluoroacetate dehalogenase
4opm	24,5	2,3	16	Putative lipase
3r3u	24,3	2,8	20	Fluoroacetate dehalogenase
1ehy	24,3	3,1	18	Soluble epoxide hydrolase
4ns4	24,2	2,5	19	Alpha/beta hydrolase fold protein
4nvr	24,1	3	18	Putative acyltranferase
2xua	24	2,5	16	3-oxoadiapate enol-lactonase
3p2m	24	2,9	22	Possible hydrolase

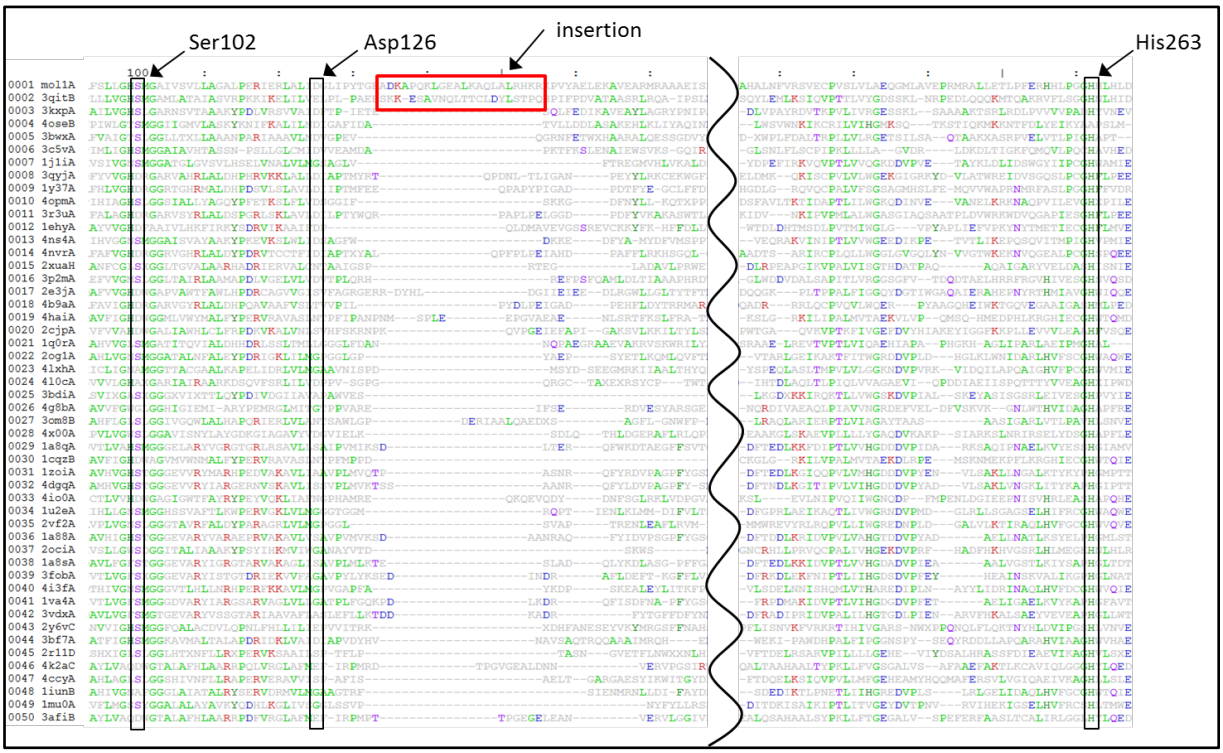


Figure 49: Sequence alignment of DALI search results reveal a rare insertion of 22 amino acids in PA1622.

This leads to a substantially longer lid domain with the exclusive formation of the protruding helix, which renders the fold of both unique. Involved in key interactions within the dimeric structure this helix is essentially responsible, supported by a specific anchoring arginine (Arg186), for locking the lid in a permanent open form conformation. Stabilizing contacts of the core domain with the lid of the other subunit locks its position on the other side. This conformation was exclusively reported for only one structure before (Gehret et al., 2011).

5.2. Structural analysis of PA1623

5.2.1. Sequence analysis

The protein sequence of PA1623 was taken from the Pseudomonas Genome Database (PGD) (Winsor et al., 2011) and analyzed using bioinformatics. SignalP (Bendtsen et al., 2004; Petersen et al., 2011) determined the protein cellular location as cytosolic (figure 50).

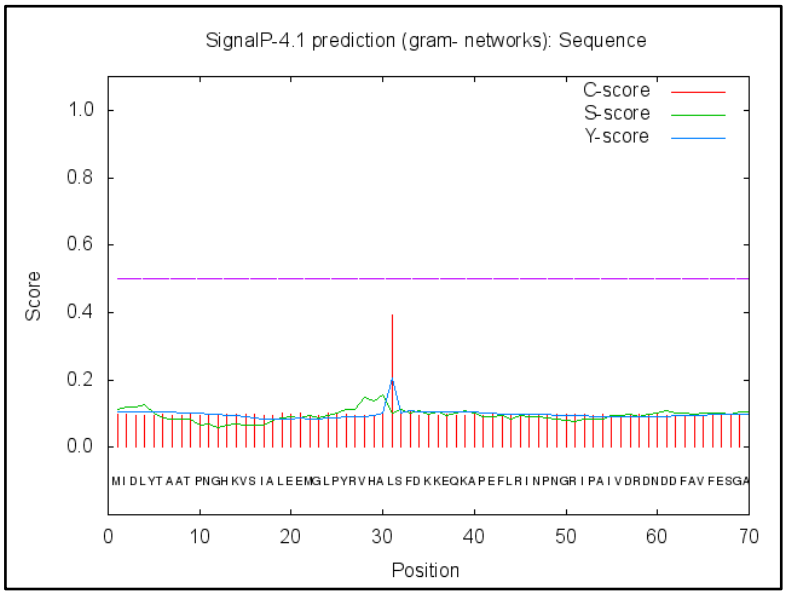


Figure 50: Prediction of cellular location of PA1623 as cytosolic (Petersen et al., 2011)

The protein has a calculated molecular mass of 24.9 kDa with an isoelectric point of 5.85 (Wilkins et al., 1999). A BLAST search (Altschul et al., 1997) indicated a conservation among *Pseudomonas species* with 32 putative orthologs, which is in agreement to automated annotation in the PGD, classifying PA1623 as 'conserved hypothetical protein'. Protein domain analysis (Marchler-Bauer et al., 2011) indicated a structure containing two conserved domains. The N-terminal one is similar to thioredoxin-like proteins while the C-terminal domain is comparable to the GST-C superfamily fold. Structural prediction carried out using the PHYRE server (Kelley & Sternberg, 2009) resulted in a possible GST-like fold protein but a score of 25 did not indicate certainty. SMART search was carried out to identify PA1623's role in functional context (Letunic et al., 2014), suggesting that it is possibly involved in glutathione metabolism (KEGG map00480) or metabolism of xenobiotics by cytochrome P450 (KEGG map00980) (Kanehisa et al., 2014). Combining all

pieces of information gathered by bioinformatics led to the conclusion that PA1623 most likely belongs to the cytosolic glutathione-S-transferase superfamily, indicated by a possible glutathione binding site and protein dimerization interface revealed by CDD analysis (figure 51) (Marchler-Bauer et al., 2011).

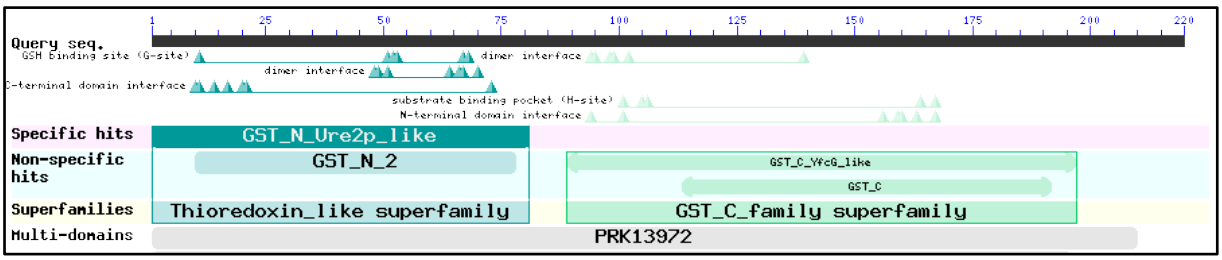


Figure 51: Bioinformatic analysis of conserved domains using the 'Conserved Domain Database' (Marchler-Bauer et al., 2011)

Querying the *Protein Data Bank* (Berman et al., 2000) revealed related proteins with potential structural similarity to PA1623. The 10 best ranked hits are shown in table 31.

Table 31: Results from BLAST search against the Protein Data Bank. The asterisk is explained in the following paragraph.

Protein	Origin	Sequence identity (%)	Query coverage (%)	E-value	Pdb code
GST	*Pseudomonas aeruginosa PACS2*	99	100	1e-161	4ECI (*)
GST	*Burkholderia graminis*	52	91	2e-63	4MF5
GST	*Xenorhabdus nematophila*	47	88	2e-61	4L8E
GST	*Bradyrhizobium sp*	50	91	9e-59	4MF7
GST	*Pseudomonas fluorescens Pf-5*	47	91	8e-57	4IKH
Oxidoreductase	*Escherichia coli K-12*	46	95	7e-56	3GX0
GST	*Pseudomonas putida*	47	92	4e-55	4NAX
GST	*Phanerochaete chrysosporium*	46	88	8e-48	4F0B
GST	*Streptococcus sanguinis Sk36*	41	91	2e-41	4MZW
GST	*Escherichia coli*	39	91	9e-41	3C8E
GST	*Streptococcus pneumoniae atcc 700669*	38	92	1e-37	4PUA
GST	*Lodderomyces elongisporus*	35	93	5e-36	4IVF
GST	*Saccharomyces cerevisiae*	36	86	7e-36	1JZR

5.2.2. Note of status

While this study was carried out the structure of PA1623 was deposited to the *PDB* on 18th of April 2012 (pdb: 4ECI) by EFI (Gerlt et al., 2011) without any related publication, yet. Therefore this structure is the best matching hit with 100% sequence identity in table 31, marked with an asterisk.

5.2.3. Short background to Glutathione-S-transferase

The promiscuous superfamily of glutathione-S-transferase (E.C. 2.5.1.18) consists of three distinct classes, membrane-associated, mitochondrial and cytosolic members. Their main function is detoxification by conjugation of electrophilic endogenous or xenobiotic compounds to glutathione rendering them less toxic and more polar. Apart from this, multifunctional cytosolic members were reported that are involved in the reduction of oxidized cellular macromolecules, function in the biosynthesis of metabolites (Oakley, 2005) and are involved in signaling pathways (Laborde, 2010; Davis et al., 2011). The dimeric proteins specifically bind the tripeptide glutathione within their N-terminal domain. The second domain comprises a rather variable amino acid pattern rendering it structurally rather flexible that is necessary to allow diverging substrate specificity for every GST.

5.2.4. Structure solution of PA1623

Crystals of native PA1623 were obtained from a mother liquor composition of 0.1 M Na/K tartrate, 0.1 M Bis/Tris pH 6.7 and 18% PEG 3350 using a protein concentration of 15 mg/ml. Trials using molecular replacement with protein structures available in early 2010 failed. Hence, the structure of PA1623 was solved by SAD phasing, using the anomalous signal of the 7 L-selenomethionine residues incorporated in the polypeptide chain. The labeled protein crystallized in space group $P2_12_12$ with 2 molecules in the asymmetric unit. Heavy atom sites were located by SHELX (Sheldrick, 2010) and used as additional input file to AutoSol used for structure solution and initial model building (Zwart et al., 2008). Obtained phases were then transferred to a native apo data set collected to a higher resolution.

Crystals of the complex were obtained by co-crystallization with 10 mM glutathione from the same mother liquor composition as the apo crystals but belonged to space group $P2_12_12_1$. Other than in the apo structure 4 monomers were identified in the asymmetric unit. Data collection statistics are summarized in table 32.

Table 32: Data collection statistics for PA1623 apo and PA1623 bound to GSSG

Data collection	PA1623 SeMet [α]	PA1623 apo [ß]	PA1623 - GSSG [ß]
Detector	Pilatus 6M	Pilatus 6M	Pilatus 6M
Wavelength (Å)	0.9779	0.91841	0.91841
Resolution range (Å)	48.7 – 2.0 (2.05 – 2.0)	47 – 1.65 (1.68 – 1.65)	45 – 1.30 (1.32 – 1.30)
Space group	$P2_12_12$	$P2_12_12$	$P2_12_12_1$
Unit cell parameters (Å)	87.5 97.5 51.2	87.1 96.6 50.9	84.5 96.6 100.8
Total No. of measured reflections	193896 (14991)	212970 (11214)	859876 (42087)
Unique reflections	30352 (2233)	52344 (2740)	201735 (9969)
Multiplicity	6.4 (6.7)	4.1 (4.1)	4.3 (4.2)
ano. Multiplicity	3.2 (3.4)	--	--
Mean I/σ(I)	15.1 (2.6)	17.5 (2.5)	10.7 (2.6)
Completeness (%)	99.6 (95.1)	99.8 (99.8)	99.7 (100)
ano. Completeness (%)	96.8 (93.1)	--	--
Average Mosaicity [°]	0.134	0.201	0.105
R_{merge} (%)	5.9 (55.3)	6.8 (63.1)	8.2 (54)
R_{meas} (%)	7 (65.5)	7.8 (72.5)	9.4 (61.8)
R_{pim} (%)	3.2 (25.2)	3.8 (35.3)	4.4 (29.5)
CC(1/2) [°]	0998 (0.932)	0.999 (0.80)	0.997 (0.792)
Matthews coeff.	2.19 (2 monomers)	2.15 (2 monomers)	2.07 (4 monomers)

[#]Data were collected from single crystal; [+]Values in parentheses refer to the highest resolution shell. [α]Data collected at SLS, Villingen, Switzerland (Willmott et al., 2013) [ß]Data collected at BESSY II, BL14.1 (Mueller et al., 2012), [°]Mosaicity and CC(1/2) reported by XDS (Kabsch, 2010; Karplus & Diederichs, 2012), [$]Rpim = Σhkl (1/(N – 1))1/2 Σi | Ii (hkl) – < I (hkl) > | / Σhkl Σi Ii (hkl), (Weiss, 2001).

Manual model adjustment were done by using COOT (Emsley et al., 2010) and phenix.refine (Afonine et al., 2012) was used for automated refinement and phase improvement. Refinement statistics are listed in table 33.

Table 33: Refinement statistics of PA5507

Refinement	PA1623 apo	PA1623 GSSG bound
Resolution (Å)	48 – 1.65 (1.69 – 1.65)	45 – 1.3 (1.34 – 1.3)
Number of reflections	49627	201614
R_{work} (%)	16.6	11.2 (19.4)
R_{free} (%)	20.6	14.3 (21.9)
No. atoms *	3671	7994
Protein	3316	6752
Water	345	1054
Average B-factors (Å²)	18.2	13.15
Protein	15.4	11.9
Water	31.1	27.4
R.m.s. deviations		
Bond length (Å)	0.022	0.013
Bond angle (°)	1.140	1.49
Rotamer outlier (%)	0	0
Ramachandran plot (%)		
Favored region	98.5	97
Outlier	0.24	0
Validation score		
Clashscore	2.5	2.47
MolProbity[&]	1.33	1.25
PDB entry code		

* indicates the number of non-hydrogen, non-solvent atoms; [&]As reported by MolProbity ((Chen et al., 2010), http://molprobity.biochem.duke.edu/).

5.2.5. Structure of apo PA1623

The polypeptide chain of PA1623 adopts a classical canonical glutathione-S-transferase (GST) fold (figure 52). Two identical monomers, which are rotated by 180 degree, form a stable protein dimer. The smaller N-terminal domain (86 residues) adopts the common topology of a thioredoxin-like domain amongst GST and GST-like enzymes (Katti et al., 1990; Martin, 1995).

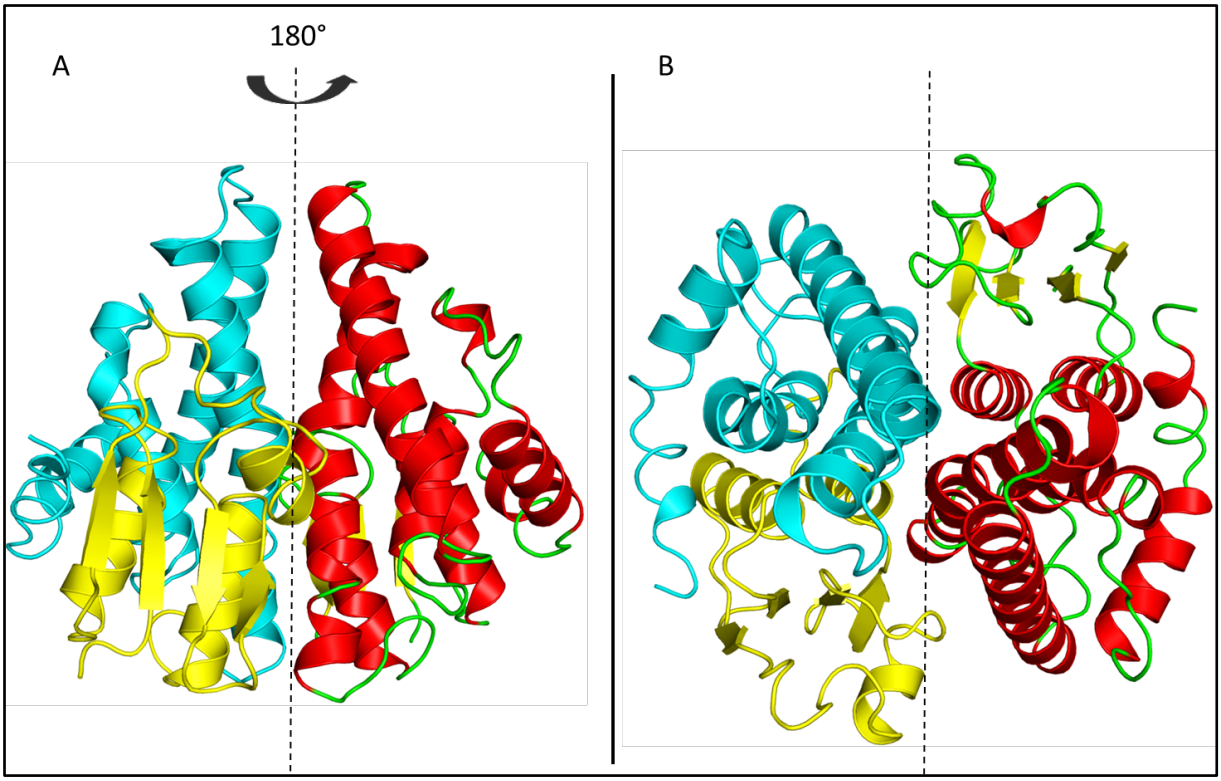


Figure 52: Apo-structure of the protein homo dimeric PA1623. Beta-sheets and alpha-helices as secondary structure elements are colored in yellow and red, respectively. Loop regions are colored in green. One chain is colored according to the domain architecture with the thioredoxin domain colored in light yellow and the C-terminal domain stained in blue.

The characteristic fold consists of four almost planar beta sheets surrounded by three alpha helices (figure 52A). Conserved amongst all members of the GST super family is a proline residue (P53) located in the loop region connecting β_2 and β_3 as alignment of 1000 protein sequences revealed (figure 53).

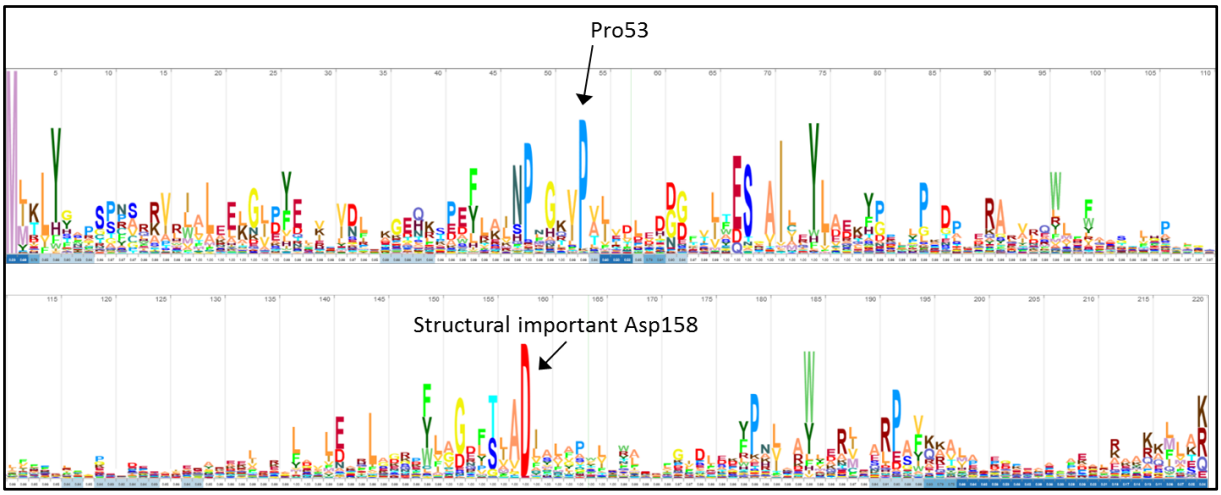


Figure 53: Sequence alignment of 1000 members of the GST super family revealed the conservation of the *cis*-proline (P53) and the structural important aspartate (D158). The total number of 110 residues is depicted within one line.

This proline residue is in the less favored *cis*-conformation and terms this loop the *cis*-Pro loop which is a catalytically important component (Allocati et al., 1999; Nathaniel et al., 2003). The proline residue (P53) of PA1623 is shown in comparison with the close homologues, table 1, *PDB*: 4MF5 (Gerlt et al., 2011) and *PDB*: 3GX0 (Wadington et al., 2009) (figure 54B).

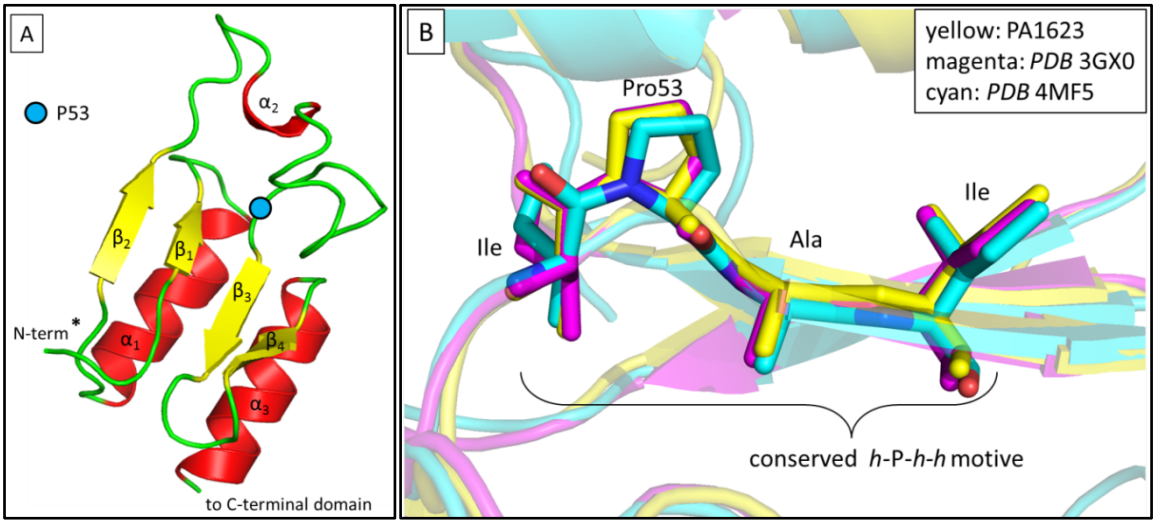


Figure 54: Cartoon representation of the N-terminal thioredoxin-like domain (A). The *cis*-proline residue which is conserved amongst members of the glutathione-S-transferase family located in a characteristic motive of *h*-P-*h*-*h* (B), where *h* is a hydrophobic residue.

High conservation of this domain already indicates a main function of it. In fact, it contributes most to the binding of the glutathione molecule as it provides large parts of

the GSH binding pocket. It is connected via a short linker containing of six amino acids (80-86) to the C-terminal domain, which is comprises an all alpha helical fold (residues 87-220). The five alpha helices wrap around each other, connected by short loop sequences. The domain begins with a long alpha helix, which cross-spans the whole structure and serves as major contributor to the protein dimerization interface. The interactions between the protein chains are mostly hydrophobic. Furthermore this domain is mainly involved in recognition, interaction as well as binding of the mostly hydrophobic substrates. Its diversity in composition introduces substrate specificity within the GST family (Wilce & Parker, 1994). A phenylalanine residue (F66) preceding from the loop extending β_2 to the C-terminal domain, acts as a key residue pointing into a 'hydrophobic lock' generated by residues of helix α_1' and α_2' of the other subunit (Reinemer et al., 1991; Sinning et al., 1993; Xiao et al., 1996; Rossjohn et al., 1998a) (figure 55B). Mutations introduced at this position (F66) cause limitations in dimerization and protein stability.

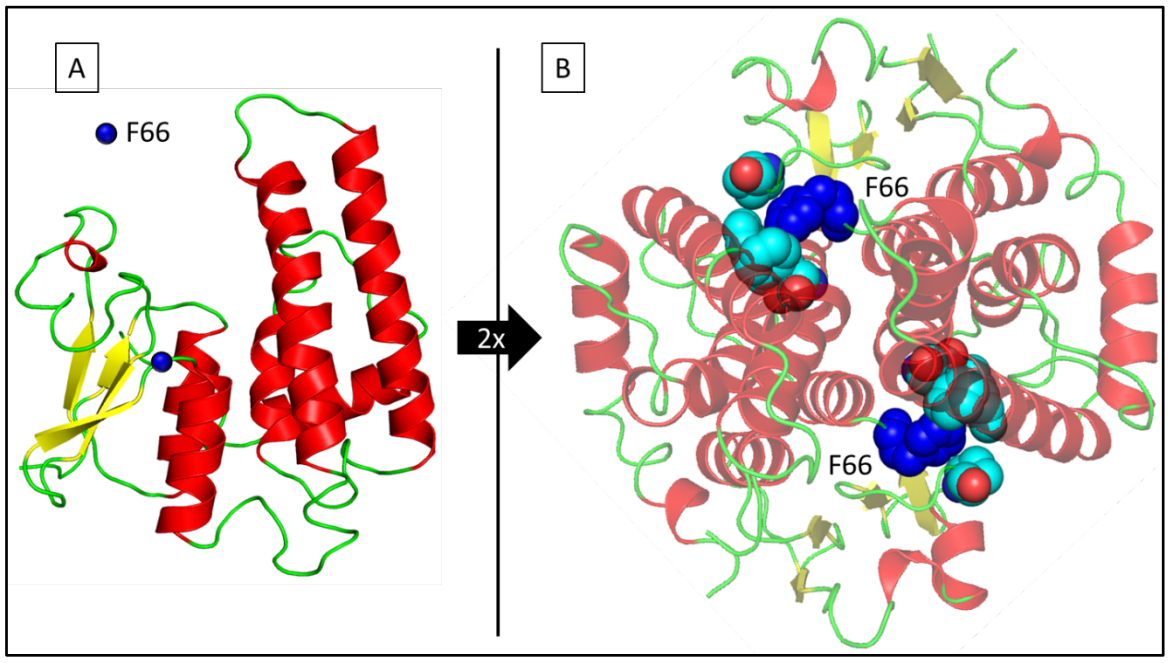


Figure 55: PA1623 comprises two domains, the N-terminal thioredoxin-like domain and alpha-helical C-terminal domain (A). Its dimer interface features conserved residues acting as 'key-lock' (F66) which serves as one of the major contributors of dimerization, highlighted in dark blue and cyan (B).

Both protein chains were modeled continuously into the electron density until amino acid 202. The last eighteen C-terminal amino acids could not be fit due the lack of electron density. Furthermore, during refinement extra density was located in the binding region of GSH. This could not be interpreted by the protein chain and was hence fitted by a

tartrate molecule originating from the crystallization condition. It occupies almost the same position as the gamma-glutamyl moiety is located in comparable GST proteins (figure 56).

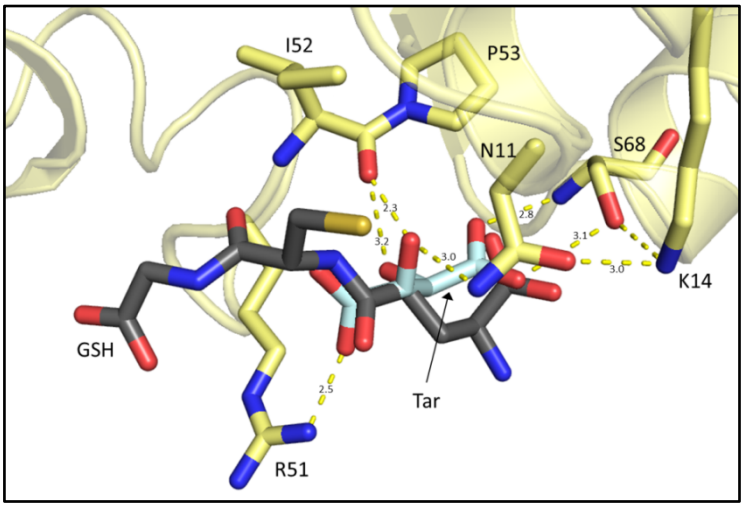


Figure 56: A tartrate molecule, found in the structure is occupying the binding site of GSH. The stabilizing hydrogen bond network is indicated by yellow dashed lines labeled with the respective distances in Ångström.

5.2.6. Structure of PA1623 bound to glutathione

As a member of the glutathione-S-transferase family, PA1623 is expected to bind glutathione (GSH). Binding affinities for both reduced (GSH) and oxidized glutathione (GSSG) were determined using fluorescence titration (Schmidpeter & Schmid, 2014) (figure 57).

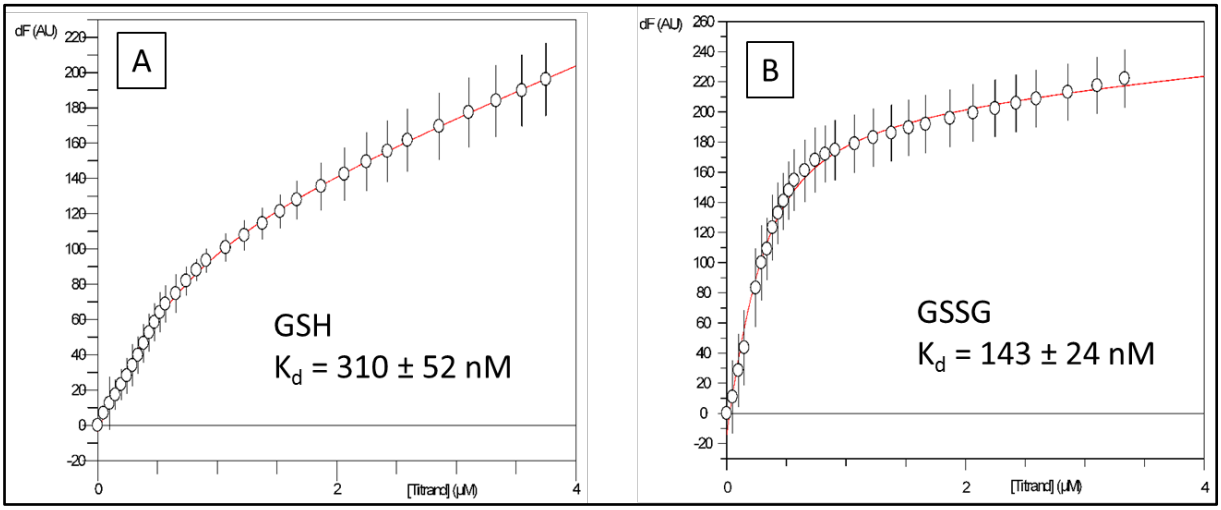


Figure 57: Fluorescent titration of GSH (A) and GSSG (B) to determine binding affinity to both to PA1623.

Binding affinities were determined for reduced glutathione and oxidized glutathione to 310 nM and 143 nM, respectively. This work was carried out during a bachelor thesis (Ludewig, 2012).

Subsequently, the ligand structure was solved by molecular replacement carried out with Phaser (McCoy et al., 2007) employing the apo structure as search model. The software placed four molecules in the asymmetric unit, forming two dimers. The structure was refined to a resolution of 1.30 Å with values of 11% and 14% for R_{work} and R_{free}, respectively (table 32 and 33). During refinement clear additional electron density was observed in all four monomers. Surprisingly, this could not be interpreted by only one glutathione molecule but two molecules per monomer fit the density.

The glutathione molecules, bound in an extended manner are stabilized by a large number of hydrogen bonds. Main contributing parts are the loop region connecting β_2 and β_3 including α_2 of the N-terminal domain mediates contact with the first glutathione molecule (figure 58). The side chains of residue K14, E67 and S68 as well as water molecules bind the terminus of gamma-glutamyl-moiety. The backbone carbonyl of P53 and the amide of I52 keep the cysteinyl part in place, while Q38 and an extended hydrogen bond network spun up by R51, K39 also involving the second glutathione molecule to compensate the negative charge of the free glycinyl moiety facing the solvent.

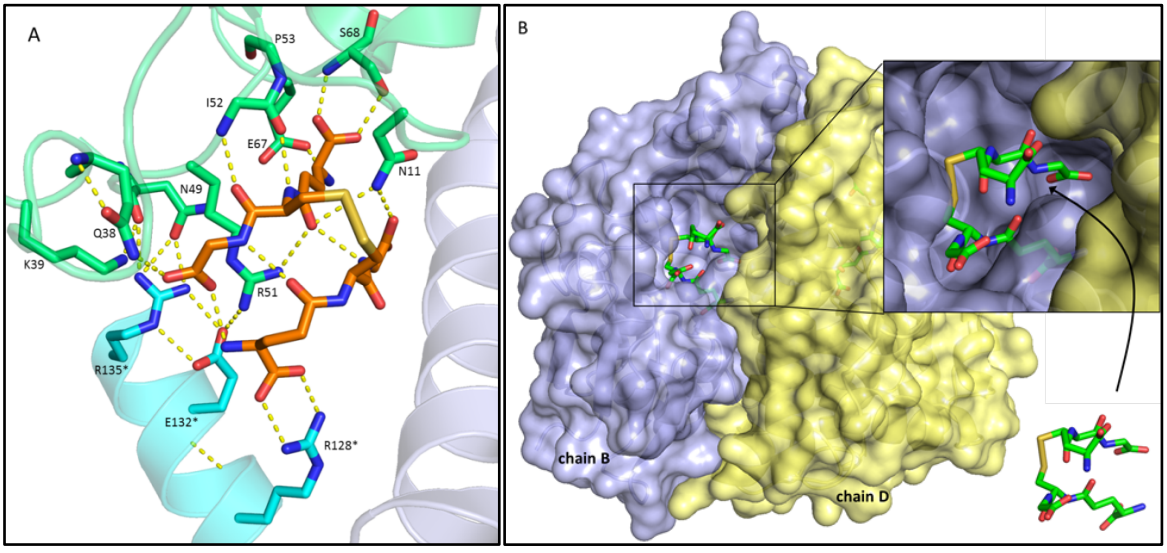


Figure 58: Binding mode of GSSG to PA1623 (A). Contribution from one amino acid chain is colored in light green (N-terminal domain) and grew (C-terminal domain). Side chains that are located on the second subunit are marked with an asterisk (cyan). GSSG binds to a crevice at the interface between both subunits of the dimer (B).

The second GSH is bound in an antiparallel manner. The negatively charged gamma-glutamyl moiety is stabilized by a salt bridge to an arginine residue, located on the α_1' helix of the other subunit. The protein-inside facing glycinyl part is hold into place by N11 and three water molecules. In addition to water molecules that fund the majority mediating contact between this glutathione molecule and the protein backbone, a few hydrogen bonds between the GSH molecules are observed. The two ligands are bound in a crevice at the interface generated by both subunits of the protein dimer (figure 58B). This renders the enzyme not a regular glutathione transferase with activity in stage II of cell detoxification but rather makes a different function likely. Despite the fact that PA1623 was crystallized with reduced GSH the oxidized form, GSSG was bound to it. This suggests that the protein might be possess disulfide-oxido-reductase activity similar to thioredoxin (Aslund et al., 1997) or glutaredoxin (Vlamis-Gardikas et al., 1997). Recently, Armstrong and coworkers reported an oxidoreductase mechanism, related to that of glutaredoxin 2 for two homologous *E. coli* proteins (Wadington et al., 2009; Stourman et al., 2011) (figure 59A). A superposition of both with the PA1623-GSSG complex revealed a remarkable structural similarity with rmsd values of 0.55 Å and 0.72 Å, respectively. However, YfgU (*PDB*: 3C8E) comprises N- and C-terminal extensions which are not present in the *Pseudomonas* protein but YfcG (*PDB*: 3GX0) shows an almost identical fold (figure 59B).

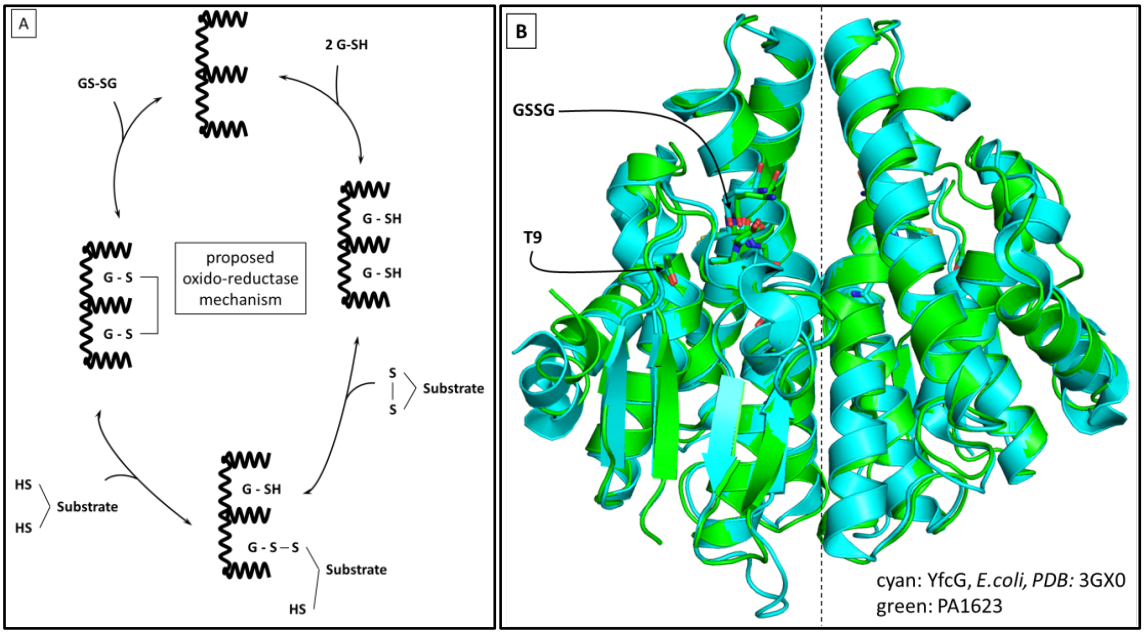


Figure 59: Proposed mechanism of PA1623 following an oxido-reductase mechanism (A) (Stourman et al., 2011) involving two glutathione monomers bound to the active site of the protein. The superposition with glutathione bound YfcG PA1623 reveals an almost identical fold with an r.m.s.d. of 0.59 Å.

Following this mechanism PA1623 was tested with the model substrates di-(2-hydroxyethyl)-disulfid (HED) and oxidized DTT using a coupled assay, monitoring the depletion of NADPH at 340 nm (figure 60).

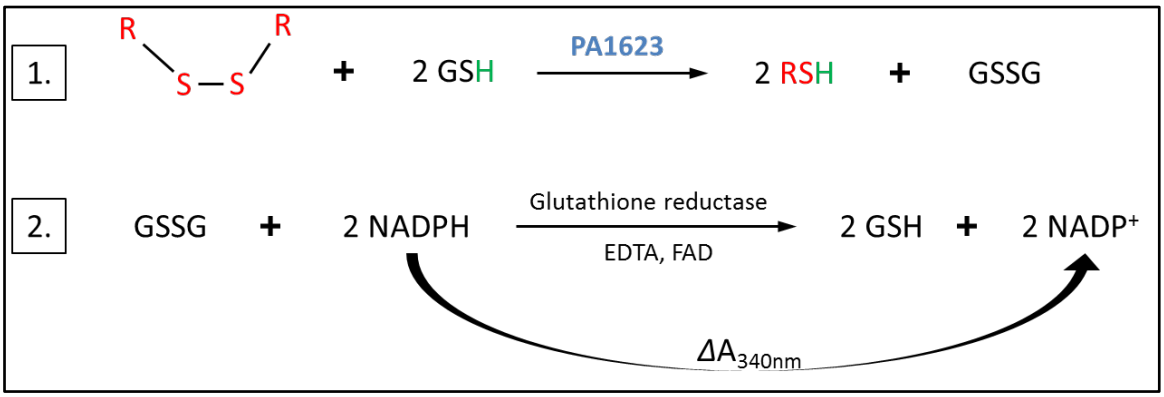


Figure 60: Coupled colorimetric enzyme reaction to assay a putative oxidoreductase activity.

This proposed mechanism however could not be verified for PA1623 using either of the substrates. The positive control experiment was carried out by spiking GSSG to the reaction. Concluding, either PA1623 does not convert these substrates or does not follow this oxidoreductase mechanism in general.

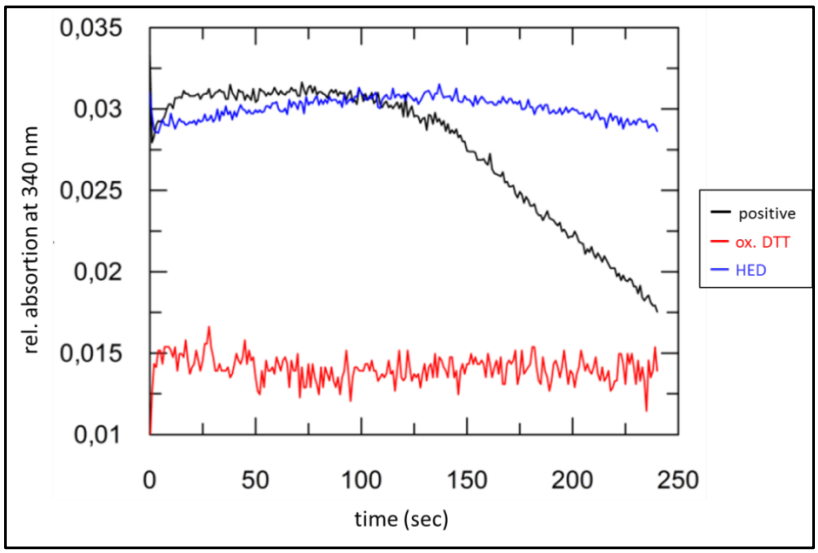


Figure 61: Oxidoreductase assay: GSSG was used in the positive control (black). The reaction curves in blue and red correspond to the substrates di-(2-hydroxyethyl)-disulfid (HED) and oxidized DTT (Braun, 2013).

Hypothetical transferase activity was tested using the general GST-substrate CDNB. These experiments were conducted during a bachelor thesis and revealed that PA1623 does not

accept CDNB as a substrate. Neither with a concentration dependent assay nor during a long term measurement any significant conversion rate could be detected (Ludewig, 2012) (figure 62).

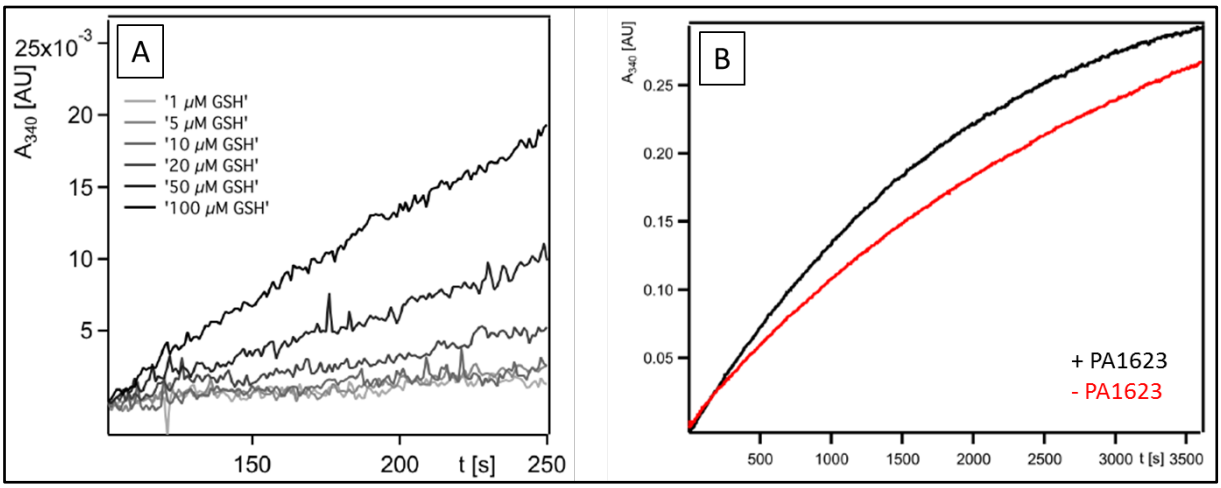


Figure 62: CDNB as a general substrate is not a substrate for PA1623 as shown by a photometric assay (Ludewig, 2012).

5.2.7. The C-terminus generates an access tunnel to GSSG

Other than in the apo structure, the C-terminus could be resolved to a further extend in the complex structure, which is clamped between the small N-terminal domain and the large cross spanning helix of one monomer (figure 63A).

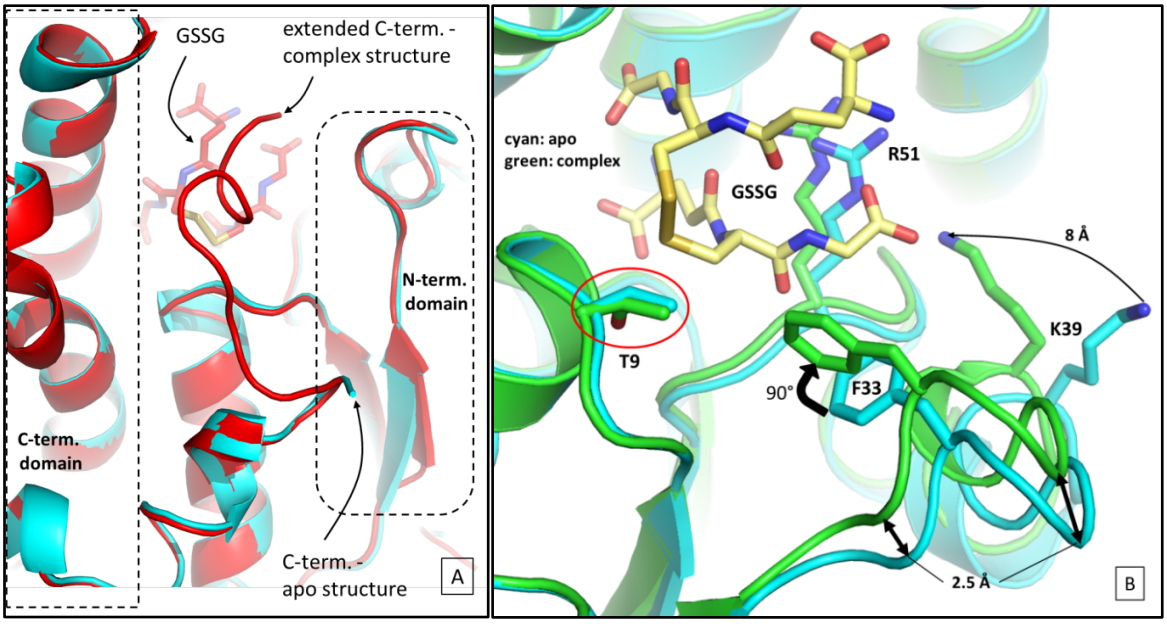


Figure 63: The resolved C-terminus of the complex structure inserts between the two domains of the protein monomer (A) and the cleft of the ligand binding site is narrowed upon glutathione binding (B).

The C-terminal fold generates a tunnel on top of the binding cavity and thereby generates a direct way to access while an arginine residue (R51) acting like a finger keeps the ligand in its position (figure 64A and 64C). In particular, this motive prevents access to the active thiol residues and shields it from the environment. An extended hydrogen bond network supported by water molecules allows directly access to the tripeptide only via the negatively charged tunnel of about 10 Å lengths (figure 64B). The tunnel originates from a flat surface patch rather neutral in charge (figure 64D). The patch is surrounded by negative charges and one positively charged arginine residue (figure 64D).

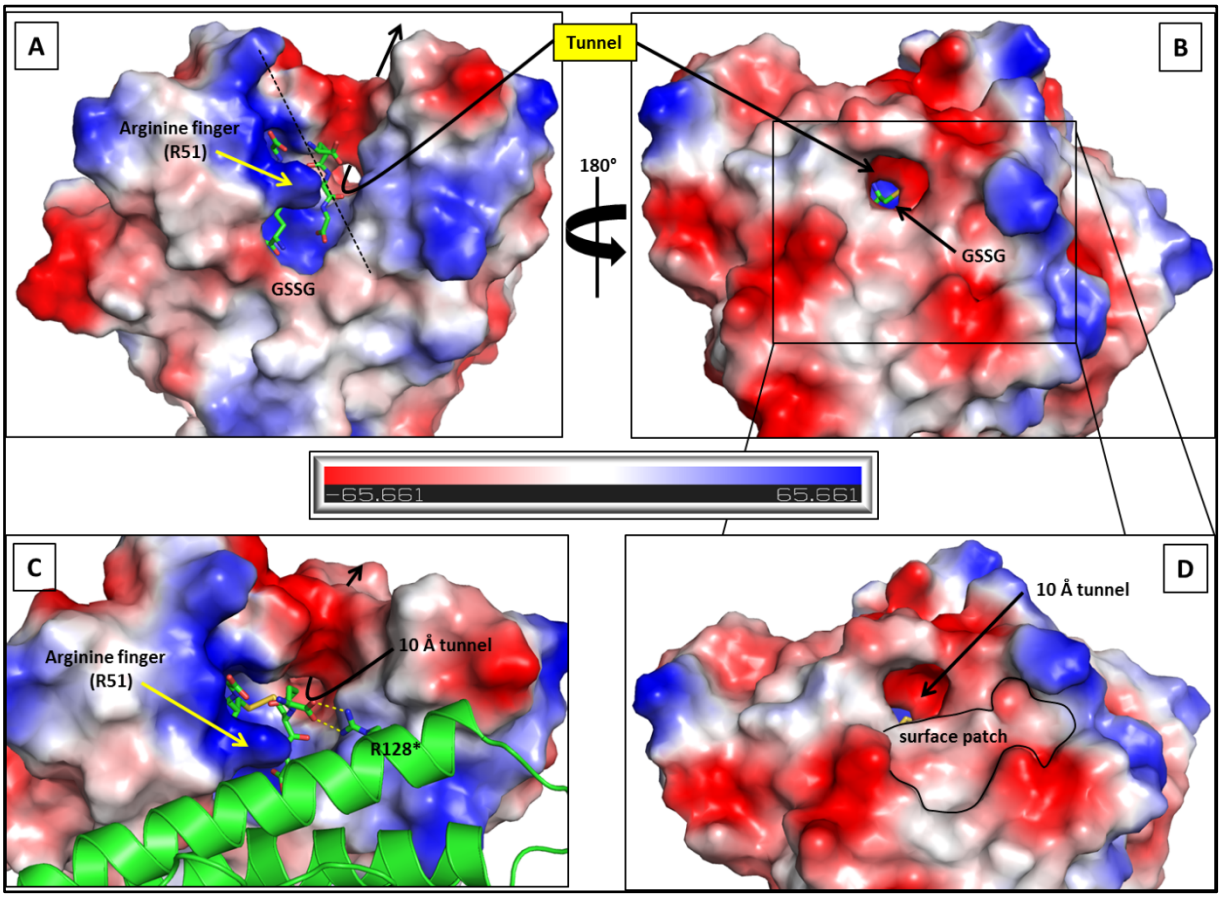


Figure 64: A tunnel is generated by the C-terminus fold of the PA1623. Via this tunnel the thiol moiety of GSSG can be accessed.

In addition, a loop region narrows the cleft in which GSSG is located by 2.5 Å upon ligand binding (figure 14B). Phenylalanine 33 undergoes a 90 degree turn and inserts between the ligand and the glutathione-binding-loop. The sidechain of lysine 39 moves about 8 Å and reorients towards the ligand, forcing the displacement of arginine 51 (figure 63B).

5.2.8. Structure of PA1623 in context

Knowledge about structure and function of GSTs was mostly gained in the past by investigations carried out on animal, human and plant representatives of the family. Only in recent years along with wide genome sequencing, the focus of GST-research was extended to bacterial members. Hence, a wide diversity in GSTs was identified and discovered iso-forms were grouped into 16 different protein classes, historically named by Greek letters: α, β, δ, κ, μ, λ, ε, ω, φ, π, σ, θ, τ, ζ and recently ν and η class. The classification criteria were originally based on structural similarities and substrate specificity (Mannervik et al., 1985) but changed to sequence similarity nowadays (Kong et al., 2003).

During catalysis mechanism, which is still not fully understood, the abstraction of a proton from GSH yielding the nucleophilic GS^- anion marks an important step. This is realized by a polar residue, located above the tripeptide bound to the protein. Due to the importance of this position the residue type has been used to classify GSTs into the classes mentioned above. Until today six different classes of soluble GSTs are known to exist in bacteria, namely the β, χ, θ, ζ, ν and η class (Allocati et al., 2009; Stourman et al., 2011; Skopelitou et al., 2012). According to sequence alignments and structural comparison it has been attempted to assign PA1623 to one of the classes.

Beta-class enzymes have a common cysteine residue at the active site and show activity to conjugate CDNB (Rossjohn et al., 1998b). In PA1623 this residue is replaced by a threonine (T9) (figure 14B), as revealed by structural comparison (data not shown). Not surprisingly, PA1623 does not conjugate CDNB (figure 13) and is therefore not a member of this class.

Chi (χ)-class enzymes, recently found in cyanobacteria, conjugate CDNB but lack the cysteine residues like PA1623. However, they feature a rare lysine residue at the catalytic position.

Theta class GSTs are found in methylotrophic bacteria allowing the bacteria to grow for example on dichloromethane (Vuilleumier et al., 2001). They use a serine residue to deprotonate and stabilize the thiolate anion, similar to the delta and epsilon class GSTs found in insects (Wang et al., 2008).

Zeta class enzymes have been identified to be involved in tyrosine as well as phenylalanine catabolism via homogentisate. They catalyze the penultimate step, an isomerase reaction with the substrates maleylacetoactetate and maleylacetone, which are converted into fumarylacetoacetate and furarylaceton, respectively (EC: 5.2.1.2) (Blackburn et al., 1998; Fernández-Cañón & Peñalva, 1998; Cornett et al., 1999). In particular, this class is characterized by a distinctive motive in the GSH-binding domain (G-site), a double serine followed by cysteine residue which are involved in the catalytic mechanism (McCarthy et al., 1996). Despite automated homology annotation suggests an isomerase activity for PA1623 following the zeta class mechanism, PA1623 does not comprise the exclusive motive present in zeta class members.

Recently, the eta (η) class (Skopelitou et al., 2012) was reported for a GST from *Agrobacterium tumefaciens*. In this protein class the polar catalytic residue is replaced by a hydrophobic loop region comprising a glycine-leucine-alanine motive. Although structurally similar, PA1623 is also not a member of this class, according to active site classification.

The nu-class (ν) enzymes have a threonine residue at the active site. Members of this class bind two glutathione molecules and were reported to have an oxidoreductase activity. Until now, only a few members have been assigned to this class, amongst them the *E.coli* proteins YghU (Stourman et al., 2011) and YfcG (Wadington et al., 2009).

In PA1623, the position of the catalytic residue is occupied by a threonine residue which would suggest a membership of the nu class. Structural comparison revealed a very similar fold of PA1623 to YfcG with an r.m.s.d. of 0.55 Å as well as a similar binding mode for two glutathione molecules. A sequence alignment of the active site containing loop showed only a difference in one amino acid, namely alanine (A8) of YfcG is replaced by a proline residue in PA1623. Overall sequence identity of PA1623 to YfcG is 46% and 39% to YghU, disregarding YghU's N-terminal extension not present in PA1623. Furthermore, YghU shows an inserted amino acid sequence of seven amino acids in the C-terminal domain resulting in an extended loop region, rendering YfcG the closest homologous protein based on sequence identity. According to the results the sequence alignments a phylogenetic tree was assigned (figure 65).

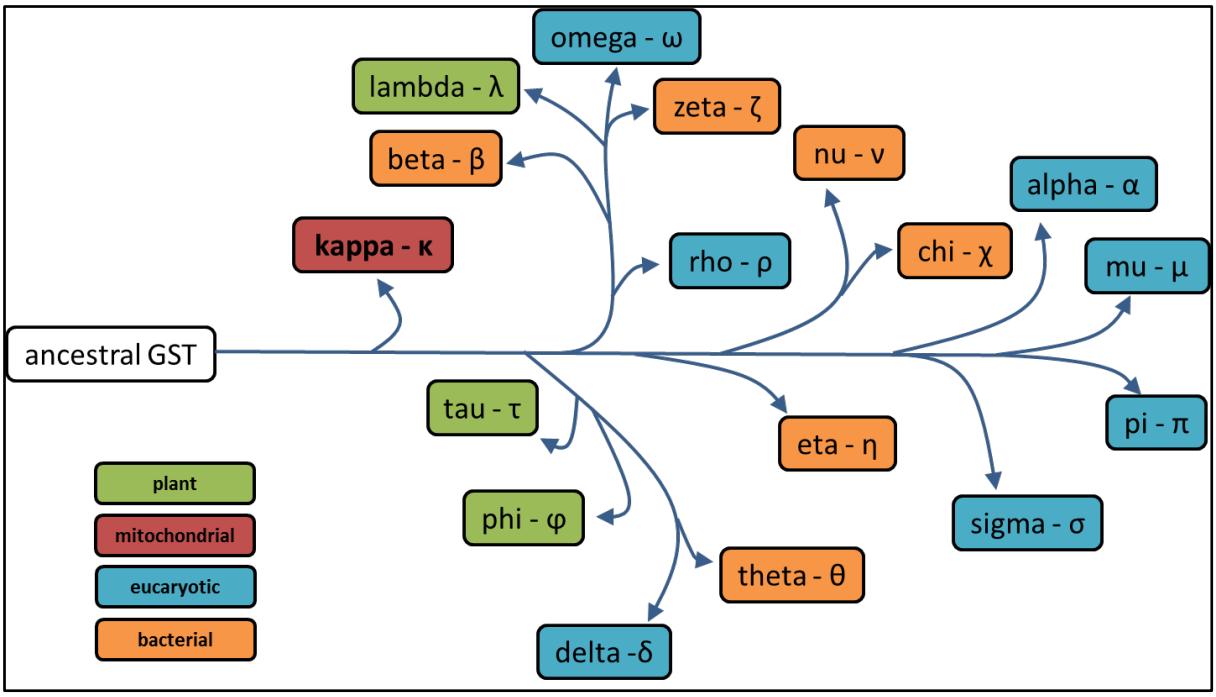


Figure 65: Phylogenetic tree resulting from sequence alignment.

In addition to sequence alignments, a DALI (Holm & Rosenström, 2010) and PDBeFold (Krissinel & Henrick, 2004) search for structural homologs was performed, using a single chain of PA1623 as query. In congruence with the sequence alignments PDBeFold also revealed the nu-class GST protein YfcG (pdb: 3GX0) as the closest homologous structure with a total score Q score of 0.83 (a match of 1 would imply an identical structure). Both search routines found as the next closest homologous structures members of the nu-class family; GSTs from *Xenorhabdus nematophila* and *Pseudomonas putida,* with r.m.s.d. values of 1.3 Å and 1.1 Å and total Q-scores of 0.77 and 0.69, respectively, both deposited by structural consortia without a related publication.

Concluding, PA1623 from *Pseudomonas aeruginosa* is a member of the family of the soluble GST-like fold proteins. The nu-class protein family is the most suited class for this protein as revealed by sequence and structural comparison. The protein binds two molecules GSH as well as GSSG with high affinity and protects the reactive thiol groups of the tripeptides by an arginine residue (R51), keeping the oxidized glutathione in place. The C-terminus of the protein generates a lid on top of the active site and at the same time generates a 10 Å long tunnel through which the thiol residue can be reached by a potential substrate.

The most homologous structure is the protein *YfcG* (Wadington et al., 2009) from *Escherichia coli* which is reported to follow an oxidoreductase mechanism. This mechanism could not be shown for PA1623 in this study and its natural substrate could not be identified.

5.3. Structural analysis of PA1624

5.3.1. Sequence analysis of PA1624

The sequence of PA1624 was taken from the Pseudomonas Genome Database (Stover et al., 2000; Winsor et al., 2011) and analysed with bioinformatics. The open reading frame codes for a 268 amino acid protein. It possesses an N-terminal signal peptide, dedicating the protein to translocation to the periplasm via the general secretory pathway (SEC). Proteins that are excreted by SEC pathway are usually translocated in an unfolded state (Tullman-Ercek et al., 2007). This indicated already a challenge for soluble recombinant protein production as expressed protein could be insoluble or is not expressed well. As signal peptides are not universally recognized amongst species, the *Pseudomonas* signal peptide might not be recognized by the *E.coli* translocase, however, to prevent any interference the protein was expressed as N-terminal truncated variant Δ_{18}PA1624, lacking its signal peptide (figure 66) (Bagos et al., 2010; Petersen et al., 2011).

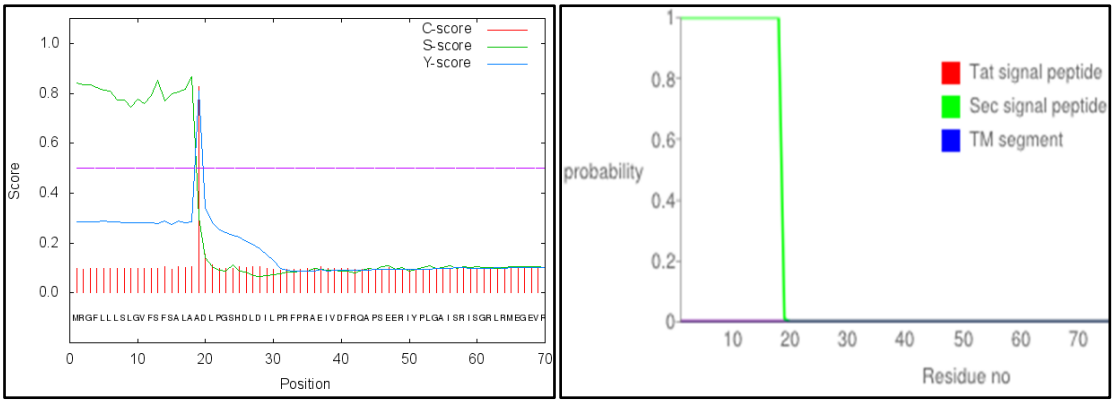


Figure 66: Leader peptide prediction of PA1624 by SignalP (left) (Petersen et al., 2011) and analysis of signal peptide type using SEC/TAT specific prediction with TMHMM (right) (Krogh et al., 2001; Möller et al., 2001).

5.3.2. PA1624 - unpredictable protein fold

A BLAST search (Altschul et al., 1997) querying the non redundant protein database reveals a conservation of PA1624 amongst *Pseudomonas species* but did not result in any significant match for other species. Analysis for potential conserved protein domains was carried out by screening more than 26 million known domains classified in 2738 protein

superfamilies for homology and similarity, thus no match was found (Sillitoe et al., 2013). The primary sequence of PA1624 was submitted to state-of-the-art structural prediction server (Kelley & Sternberg, 2009). Structural prediction of the in full length protein failed (figure 67).

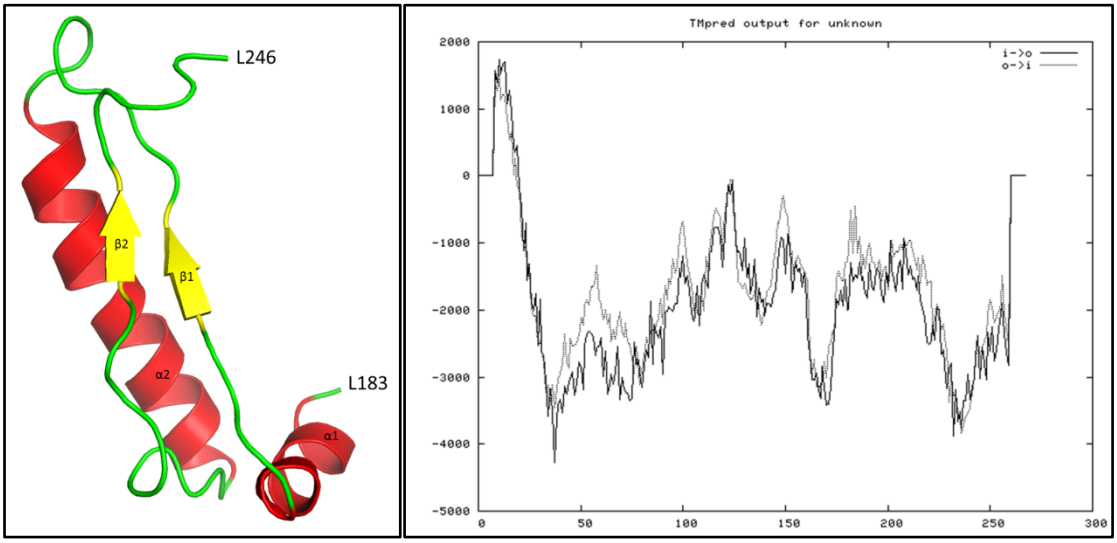


Figure 67: Structural prediction results: PHYRE result using the primary sequence of Δ18PA1624 as query (Kelley & Sternberg, 2009). Alpha helical parts are colored in red and named with α while beta strands are highlighted in yellow and named with the Greek β and unstructured loop regions are stained in green. Right: Prediction of transmembrane and membrane integral helical parts of PA1624.

The predicted structural motive covers amino acids 183-246 and comprises two alpha helices (α1 and α2) and two parallel beta sheets (β1 and β2) (figure 67). The model is based on a 65 residues stretch of the uncharacterized outer membrane protein rv0899 from *mycobacteria tuberculosis* (Li et al., 2012) with sequence coverage of 24% and identity of 18%, respectively. The model belongs to the OmpA family; PA1624 however does neither contain a signal for outer membrane integration nor any predicted transmembrane or membrane integral helical part as revealed by sequence analysis using TMHMM (Krogh et al., 2001; Möller et al., 2001) (figure 67). The only helical part that was recognized by the algorithm as transmembrane helix is the N-terminal part of the protein chain, which contains the signal peptide. The leader peptide is inserted into the membrane prior to translocation of the protein and cleaved by various signal peptidases later on (Fröbel et al., 2012).

5.3.3. Structure solution of PA1624

The protein was expressed from *p10$*, purified to 95% purity in a three step manner and concentrated to 15 mg/ml prior to crystallization.

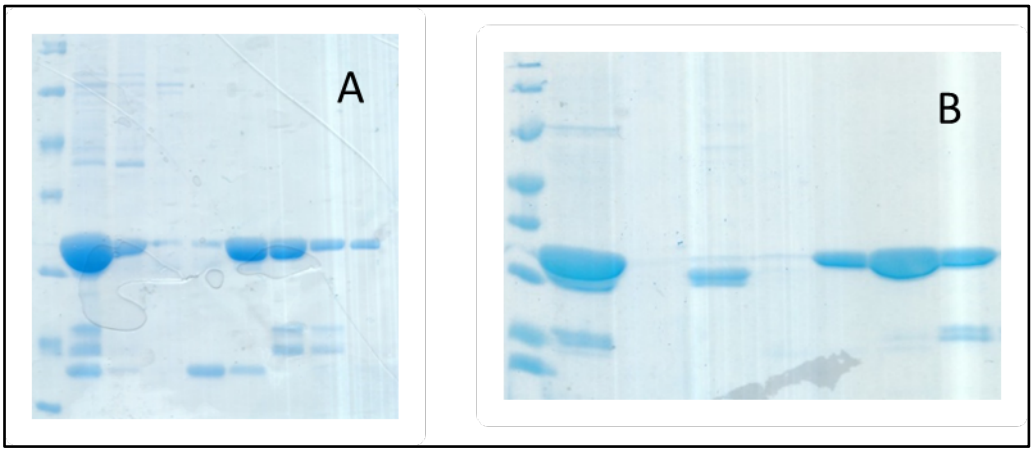


Figure 68: Purity of native Δ_{18}PA1624 (A) and selenomethionine labelled protein (B) after size exclusion chromatography.

Crystals of native Δ_{18}PA1624 were obtained from a mother liquor composition of 0.2 M sodium acetate, 0.1 M HEPES pH 7.73 and 24.5 % PEG 4000. The SeMet-labeled protein crystallized in slightly different conditions with 0.15 M sodium acetate, 0.1 M HEPES pH 7.1 and 23.3 % PEG 4000. Typical native crystals grew to a size of about 220 µm by 950 µm after 12 days and selenomethionine labeled protein crystallized in a slight different shape and size within 20 days (Figure 69). Both crystallized in space group $P2_12_12_1$ with two protein chains in the asymmetric unit and slightly different unit cell dimensions.

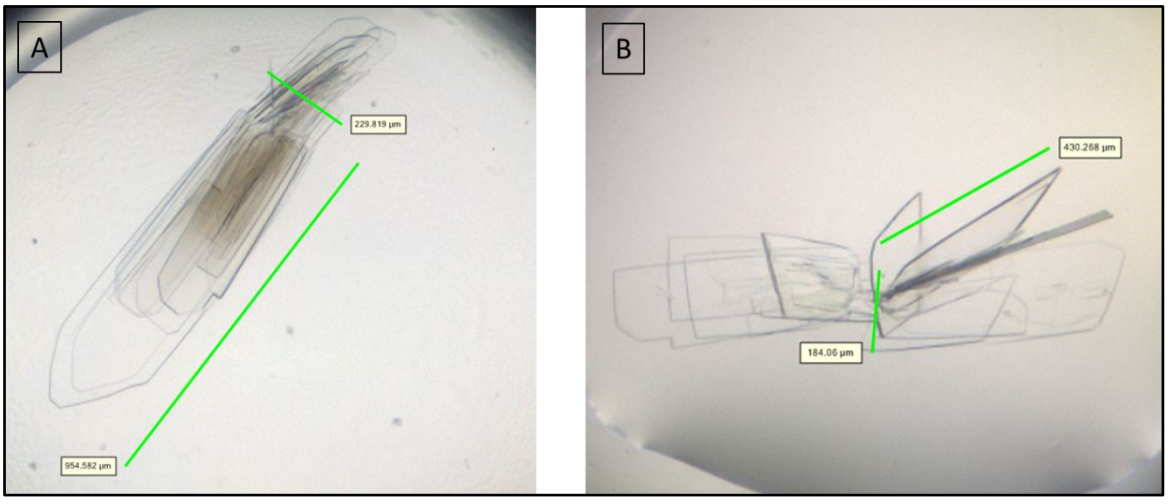


Figure 69: Native crystals of Δ_{18}PA1624 (A) and selenomethionine derivatives of Δ_{18}PA1624 (B) crystallized in slightly different conditions with slightly different shape.

The structure was solved by single anomalous diffraction of selenium collected at the absorption edge. Data were integrated by XDS (Kabsch, 2010) and the anomalous signal extracted by SHELX (Sheldrick, 2010). The located heavy atom positions were used as additional input file for AutoSol (Zwart et al., 2008; Adams et al., 2010) carrying out structure solution. Manual model adjustments were done by COOT (Emsley et al., 2010) altering with automated refinement using phenix.refine (Adams et al., 2010). Data collection and refinement statistics are shown in table 34 and 35.

Table 34: Data collection statistics

Data collection	Δ_{18}PA1624 (native) #ß	SAD-Δ_{18}PA1624 #ß*
Detector	Rayonix CCD	Pilatus 6M
Wavelength (Å)	0.918	0.979531
Resolution range (Å) +	30 – 2.4 (2.49 – 2.40)	47 – 1.95 (2.00 – 1.95)
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit cell parameters (Å; °)	54.42 58.81 163.4	53.30 59.32 158.54
Total No. of measured reflections+	78498 (8666)	3358272 (58873)
Unique reflections+	21179 (2202)	37338 (2498)
Multiplicity+	3.7 (3.9)	89.9 (23.6)
Anomalous multiplicity	-	46.5 (12.2)
Mean I/σ(I) +	4.6 (1.5)	20.3 (2.0)
Anomalous completeness (%)	-	99.6 (96.5)
Completeness (%)+	99.6 (99.8)	99.5 (96.7)
Average Mosaicity °	0.333	0.16
Matthews' coefficient	2.37	2.27
R_{merge}+ (%)	22.3 (85.2)	25.9 (210)
R_{meas}‡ (%)	26.1 (98.7)	26 (215)
$R_{pim}$$ (%)	13.4 (49.4)	2.7 (43.3)
CC(1/2)	0.976 (0.665)	0.999 (0.621)

#Dataset was collected from single crystal
*Friedel mates were treated as separate reflections
+Values in parentheses refer to the highest resolution shell.
ßData collected at MX-BL14.1 (Mueller et al., 2012)
°Mosaicity reported by XDS, (Kabsch, 2010)
‡$R_{meas} = \Sigma_{hkl} (N/(N-1))1/2 \Sigma_i | I_i (hkl) - < I (hkl) > | / \Sigma_{hkl} \Sigma_i I_i (hkl)$, where N is the number of observations of the reflection with index hkl and I_i is the intensity of its i^{th} observation.
$$R_{pim} = \Sigma_{hkl} (1/(N-1))1/2 \Sigma_i | I_i (hkl) - < I (hkl) > | / \Sigma_{hkl} \Sigma_i I_i (hkl)$$, where N is the number of observations of the reflection with index hkl (Weiss, 2001).

Table 35: Refinement statistics: Due to better data quality of SeMet-dataset the native data was left unrefined.

Refinement statistics	Seleno- Δ_{18}PA1624
Resolution (Å) [+]	47 – 1.95 (2.00 – 1.95)
$R_{work\ (\%)}$[+]	18.6 (29.1)
$R_{free\ (\%)}$[+]	22.6 (32.5)
No. of atoms	3952
Protein	3767
Water	188
Average B-factors (A^2)	
Protein	31.1
Water	29.2
R.m.s deviations	
Bond lengths (Å)	0.008
Bond angles (º)	1.007
Ramachandran plot (%)	
Favoured regions	98
Outliers	0.53
Validation score[&]	
Clashscore	0.53
MolProbity score	0.74

[+]Values in parentheses refer to the highest resolution shell.
[&]As reported by MolProbity (http://molprobity.biochem.duke.edu/ (Chen et al., 2010)

As native data were collected to a lower resolution the structure was left unrefined. The final model was refined to a resolution of 1.95 Å with a R_{work} of 18.6% and R_{free} of 22.5 %. The protein geometry is close to the ideal case with all residues in the allowed and 98% in favored region of the Ramanchandran plot. MolProbity (Davis et al., 2004), used for structural validation revealed an overall score of 0.74 (Chen et al., 2010) (table 35).

5.3.4. PA1624 Structure and homology search

The polypeptide chain folds into a two domain protein. The N-terminal domain, covering the first 175 amino acids, is connected by a linker of ten amino acids to the smaller C-terminal domain consisting of 82 amino acids. Both domains include alpha-helical (in order of primary sequence N-6-8-12-14-16-C) and beta-sheet (in order of primary sequence N-1-2-3-4-5-7-9-10-11-13-15-17-18-C) secondary structure elements (figure 70).

The first 36 amino acids do not show any significant secondary structure, however are stabilized by hydrophobic interaction with a long helix (α_1) covering amino acids 88-102 of the same domain (figure 70). Following the chain, two beta sheets (β_1 and β_2) are connected by a loop allowing a kink of 100° introduced by a proline residue (P52) at the terminus β_1. A thumb like pair of antiparallel of beta sheets (β_3, β_4) stretches out and folds back into the core. A glycine residue (G67) allows a 90° turn which separates the thumb motive from the main core structural element, an arrangement of 6 antiparallel beta sheets, clamped between the two helices and the thumb motive. First a prominent large beta sheet (β_5) twists across the domain core followed by the longest helix of the structure (α_1) which is clamped between the N-terminus as described above and the smaller C-terminal domain.

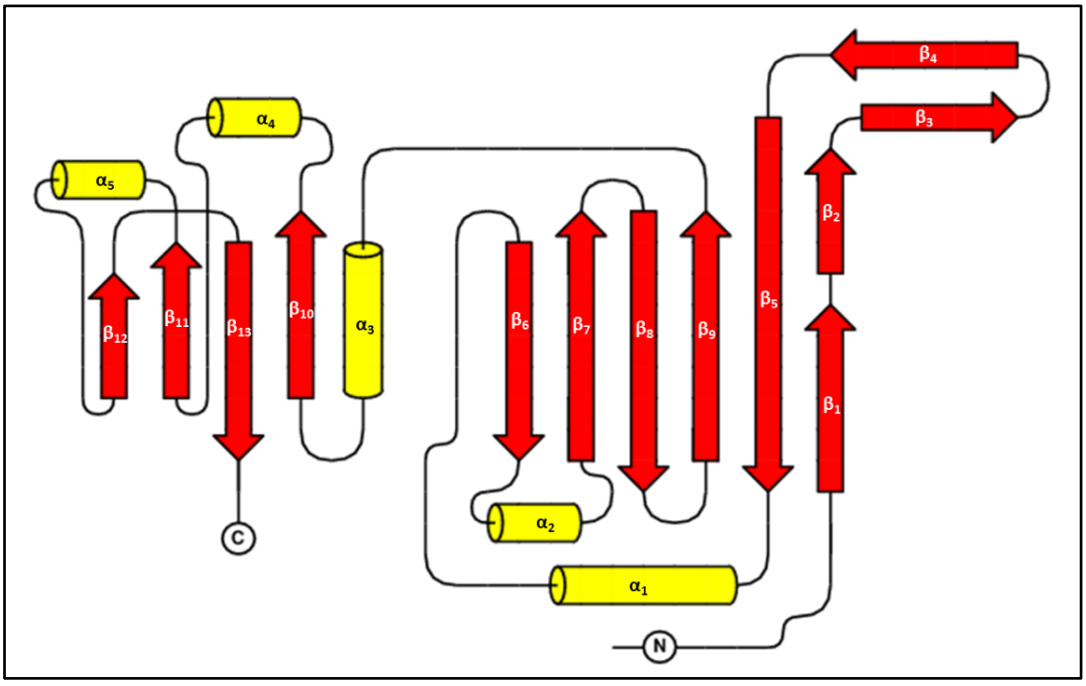


Figure 70: Overall topology of Δ_{18}PA1624: Alpha helical elements are coloured in yellow, while beta strand are stained in red. The secondary structure elements are numbered with α_x and β_x accordingly. The figure was generated with TopDraw (Bond, 2003)

The next beta sheet (β_6) is connected to a short alpha helical parts (α_2) fixed between the beta-thumb and the core beta-sheets. The core domain is completed by three antiparallel beta strands (β_7, β_8 and β_9), constructing the main part of the twisted six-stranded beta core. A linker covering amino acid D176 through S185 connects the N- and the C-terminal domain, which is an organization of four alpha-helical and four beta-sheets. The almost rectangular fold with two parallel and two antiparallel sheets on one side of the domain and the helices occupy the other three sides.

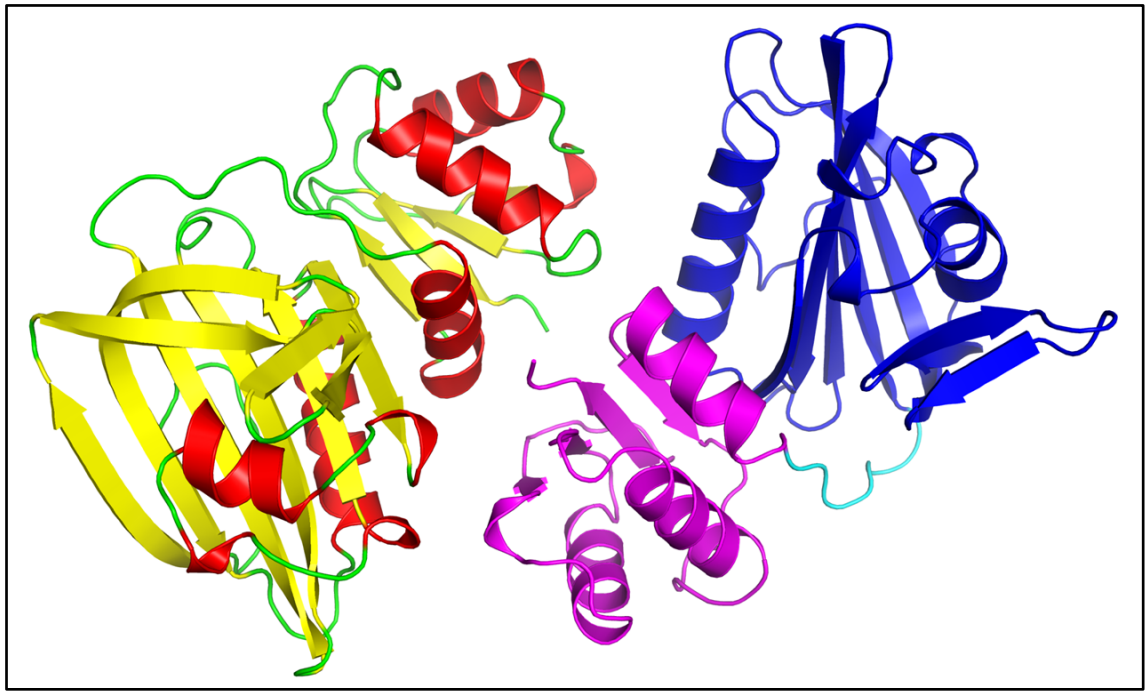


Figure 71: The structure of the crystallographic dimer of Δ_{18}PA1624. The left chain is colored according to the secondary structure elements (see also figure 70). The large N-terminal domain of the right chain is stained in blue followed by the cyan linker and the smaller C-terminal domain colored in magenta.

The N-terminal alpha helix (α_3) and the following beta strand (β_{10}) are spanning the interface between both domains. Connected by a long loop region, a double motive consisting of an alpha helix altering with a beta sheet (α_4-β_{10}-α_5-β_{11}) contributing to the outer part of the core beta blade in parallel manner follows. A long linker region bridges back to the domain terminating beta strand, β_{12} and the terminal small loop region with the terminal arginine residue, nicely resolved in the electron density. The overall arrangement of secondary structure elements is shown in figure 70 in schematic topology. The proteins overall structure is shown in figure 71 and 72.

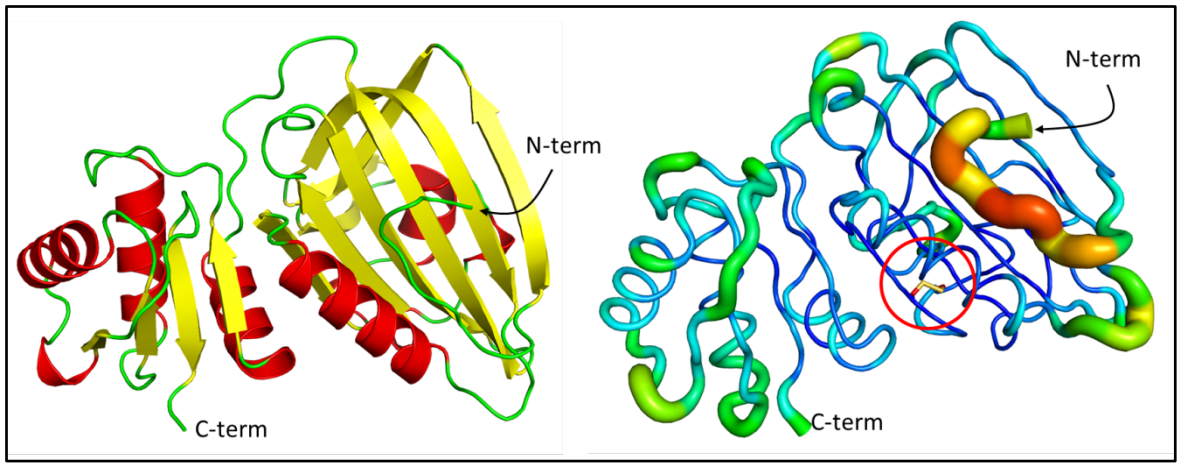


Figure 72: Structure of Δ_{18}PA1624 and B-factor analysis: Left: Overall structure of Δ_{18}PA1624, beta sheets are colored in yellow and alpha helical secondary structure elements in red. B-factor analysis revealed a very rigid (blue color) N-terminal core with rather increased (blue-green color) flexibility in the smaller C-terminal domain. The N-terminus shows the highest B-factors thus the highest flexibility indicated by the orange-red stain. The disulfide bridge is indicated by the red circle.

The monomeric protein contains one structural intramolecular disulfide bridge, involving cysteine 110 and 115, which are only 4 amino acids apart. The disulfide bridge introduces rigidity to the region by interlocking the loop region covering α_2 to the beta strand core, as indicated by a B-factor analysis (figure 72). Interestingly, disulfide bridge formation is in most gram-negative bacterial exclusive restricted to the periplasmic space, PA1624 however, was express in the reducing environment the cytoplasm. Therefore, the formation of the stabilizing bond must have happened during purification.

The two monomers within the unit cell make crystal contacts via the interaction of 15 (chain A) and 16 (chain B) residues, spread across an interface of about 400 Å². The PA1624 is a monomeric protein as indicated by PISA analysis (Krissinel & Henrick, 2007) with a surface accessible surface of 11250 Å².

To identify homologous members of Δ_{18}PA1624 the algorithm BackPhyre, implemented in the phyre search engine was used to find structural relatives (Kelley & Sternberg, 2009). The proteins structure was submitted in full length as well as the two domains separately and all available genomes were analyzed for similar structures. A blast search (Altschul et al., 1997) querying the non-redundant database excluding *Pseudomonas species* (taxid: 286) revealed that this protein seems to be only conserved amongst the originating species *Pseudomonas* as no significant sequence homology was found the NCBI database. Comparing the total scores, judged by NCBI the protein with the highest identity not

belonging to the *Pseudomonas* family is a uncharacterized hypothetical protein from *Azotobacter vinelandii* with an identify of 58% amongst 267 amino acids, while PA1624 is conserved amongst most *Pseudomonas* species.

No similar structure of structural motive was found when querying the whole *PDB* archive using searching with PDBeFold with default acceptable matches (Krissinel & Henrick, 2004; Krissinel, 2007). Even if the scoring rate was reduced to 50% acceptance no match is detected. Driving this to the edge of significance with 30% acceptance rate, a few matches are identified but with no implication, rendering PA1624 a unique structure.

PARTII

Open reading frame *PA5506-PA5509* – the *qap*-operon

5.4. Structural analysis of PA5506

5.4.1. Sequence analysis of PA5506

The sequence for PA5506 was taken from the *Pseudomonas* Genome Database (Stover et al., 2000; Winsor et al., 2011). The open reading frame (6198055 – 6198912, + strand) codes for the 32.1 kDa hypothetical protein, annotated as possible transcription factor. Bioinformatic analysis indicated a two domain protein with an N-terminal helix-turn-helix motive (HTH_6) similar to the DNA-binding domains of transcription regulators (Sørensen & Hove-Jensen, 1996) (figure 73).

Figure 73: Sequence alignment of known helix-turn-helix domains with the domain of PA5506 (Marchler-Bauer et al., 2011) and helical regions are indicated.

The C-terminal domain shows homology to phosphosugar isomerases and phosphosugar binding proteins (Teplyakov et al., 1998) (figure 74).

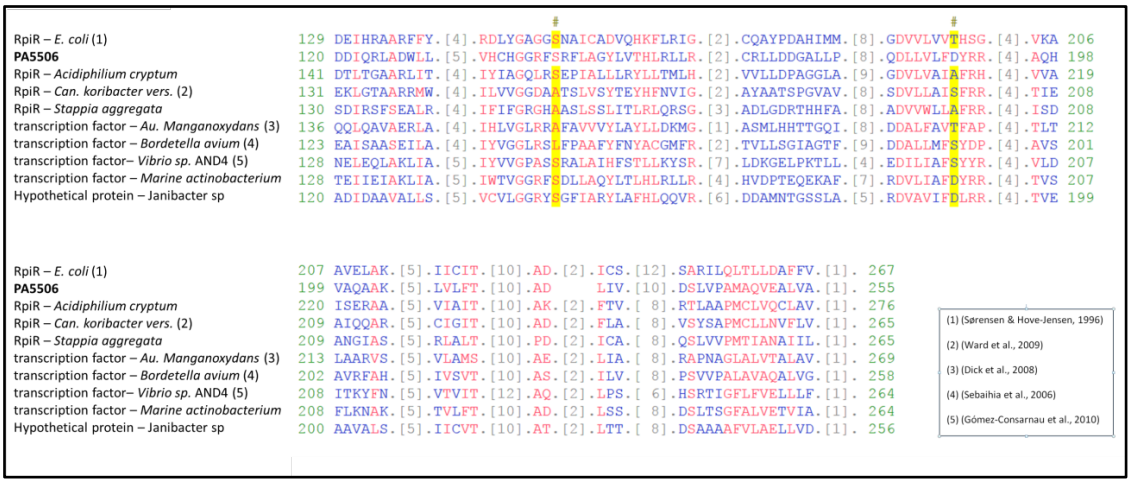


Figure 74: Sequence alignment of the C-terminal putative sugar isomerase domain of PA5506 with homologous members identified by CDD (Marchler-Bauer et al., 2013). Putative active site residues are hash marked (#).

The putative active site residues were predicted to be serine 144 and aspartic acid 188 as identified by sequence homology modelling using conserved domain analysis (Marchler-Bauer et al., 2013) (figure 75).

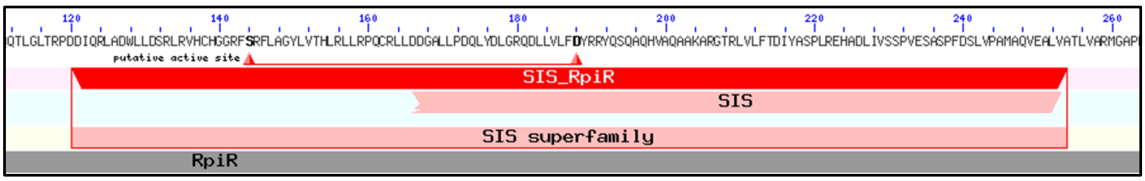


Figure 75: Predicted domain assignment. The putative active site residues of the C-terminal are marked with a triangle.

A sequence similarity search against the non-redundant protein database of NCBI carried out with HMMER (Finn et al., 2011) identified ~500 homologous sequences with an E-value < e^{-30}. PA5506 is highly conserved amongst *Pseudomonas* species. Several homologous sugar isomerases (SIS) families were identified which are either individual proteins or a domain of larger protein. Most similar proteins are members of the RpiR-like transcription regulator family. HMMER identified two structurally characterized close homologs of P5506 within the *PDB* (Bernstein et al., 1977). One hit covers the exact two domain query architecture (table 36) and the second only the SIS domain. Both are members of the transcription regulator proteins originating from *Vibrio vulnificus* (PDB: 4IVN) (Hwang et al., 2013) and *Sphaerobacter thermophilus* (PDB: 3SHO). Furthermore, two structures only matching the HTH domain were identified, originating from *Bacillus subtilis* (*PDB*: 2OF3) and *Vibrio cholera* (*PDB*: 3FXH) (Sureshan et al., 2013).

Table 36: Results from hidden Markov model search querying the *PDB* (Finn et al., 2011).

PDB-ID	Description	Origin	E-value	Coverage
4IVN	Transcription regulator	*Vibrio vulnificus*	1.1e-07	full length
3SHO	Transcription regulator, RpiR family	*Sphaerobacter thermophilus*	0.0024	SIS-domain
2OF3	HTH-type transcription regulator	*Bacillus subtilis*	0.11	HTH-domain
3FXH	Integron cassette	*Vibrio cholera*	0.7	HTH-domain

5.4.2. PA5506 – structural prediction

Results from sequence analysis indicate that PA5506 is a member of the RpiR-like transcription factor family. So it was not surprising that results from structural prediction using PHYRE revealed the already identified transcription factor (*PDB*: 4IVN) as the best homologous model for PA5506 with a TM score of 1.0 followed by the SIS domain of the RpiR-transcription regulator (PDB: 3SHO). The alignment of the twenty best models created by PHYRE (Kelley & Sternberg, 2009) shows the unique two domain structure of 4IVN including the HTH motive and the sugar isomerase domain superposed with other single SIS domains (figure 76).

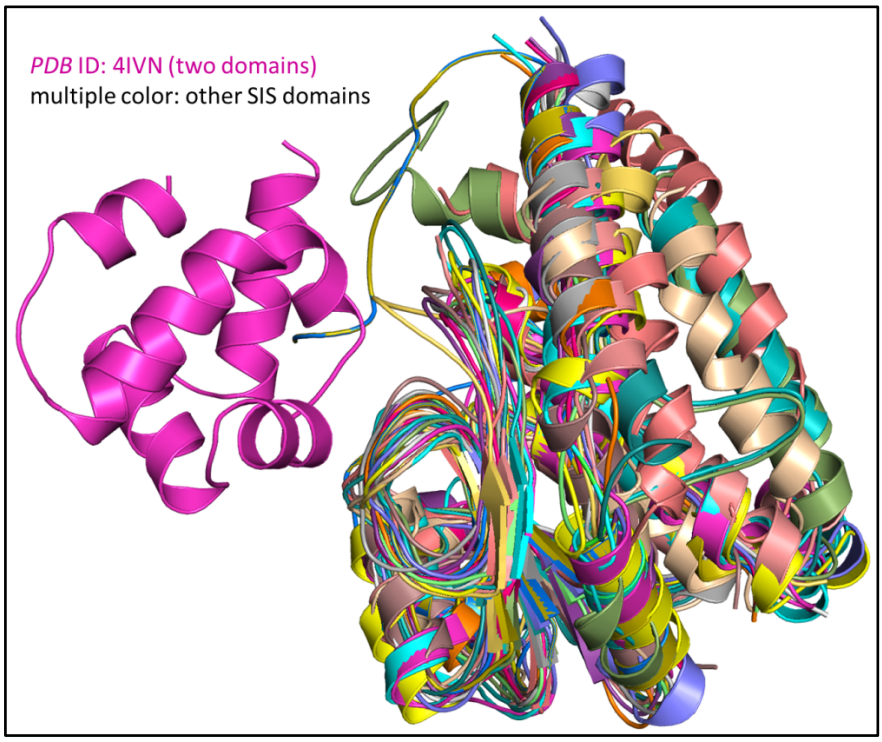


Figure 76: Superposition of predicted putative structures of PA5506 based on various *PDB* models.

5.4.3. Purification, data collection and structure solution of PA5506

Native and selenomethionine labeled protein were both expressed from *pet19mod* and purified to 95% purity according to Coomassie staining in a three step manner (figure 77).

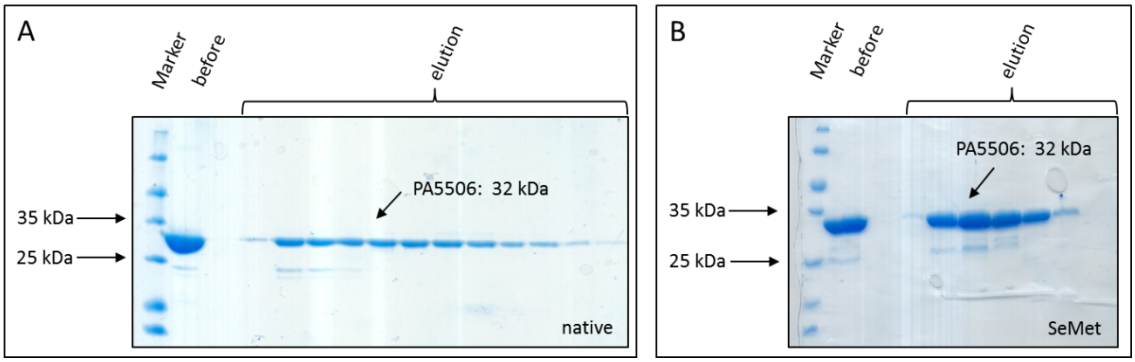


Figure 77: Purity of PA5506 after size exclusion chromatography of native (A) and selenomethionine labeled protein (B).

Labeled protein was crystallized in space groups $P2_12_12_1$ with cell dimensions of a = 67.0, b = 101.5, c = 110.0 as well as in $P4_12_12$ with a = b = 70.9, c = 157.3, both with α, β, γ = 90° (table 2 and 3). Native crystals were obtained in space group $P4_12_12$ with slightly larger cell dimensions of a = b = 71.3, c = 166.0, α, β, γ = 90° (figure 78).

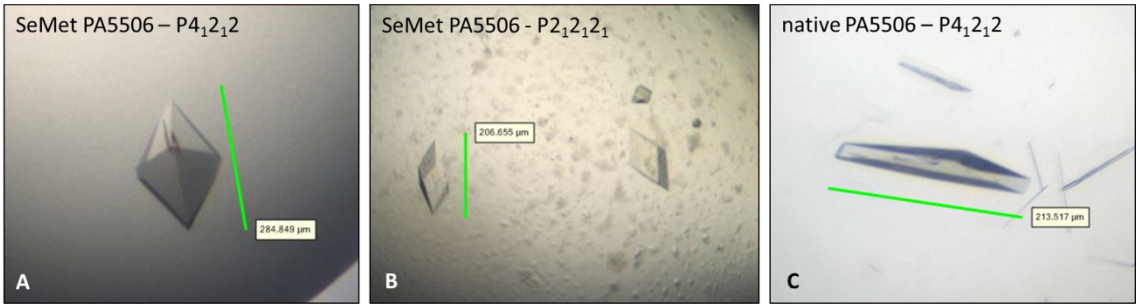


Figure 78: Typical crystals obtained from selenomethionine labeled (A, B) and native PA5506 (C). Derivative protein crystallized in (A) 5% 2-Methyl-2,4-pentanediol, 0.1 M HEPES pH 7.5, 10% PEG 6000 and in (B) 20% PEG 4000, 0.2 M potassium fluoride, 0.05 M KH_2PO_4. Native protein crystals were obtained from a mother liquor composition of 0.2 M sodium chloride, 10 % PEG 3000 and 0.1 M sodium phosphate citrate pH 4.2.

Molecular replacement either employing either the transcription regulator structure (*PDB*: 4IVN) or single sugar isomerase domains as search model was not successful. A high redundant dataset, initially collected for a sulfur SAD approach, indicated the presence of a stronger anomalous scatterer stronger than sulfur bound to the protein, possibly a metal (data not shown). This was confirmed by an extended X-ray absorption structure fluorescence scan (EXAFS) with peaks, which were unambiguously characteristic for the metal zinc (figure 79).

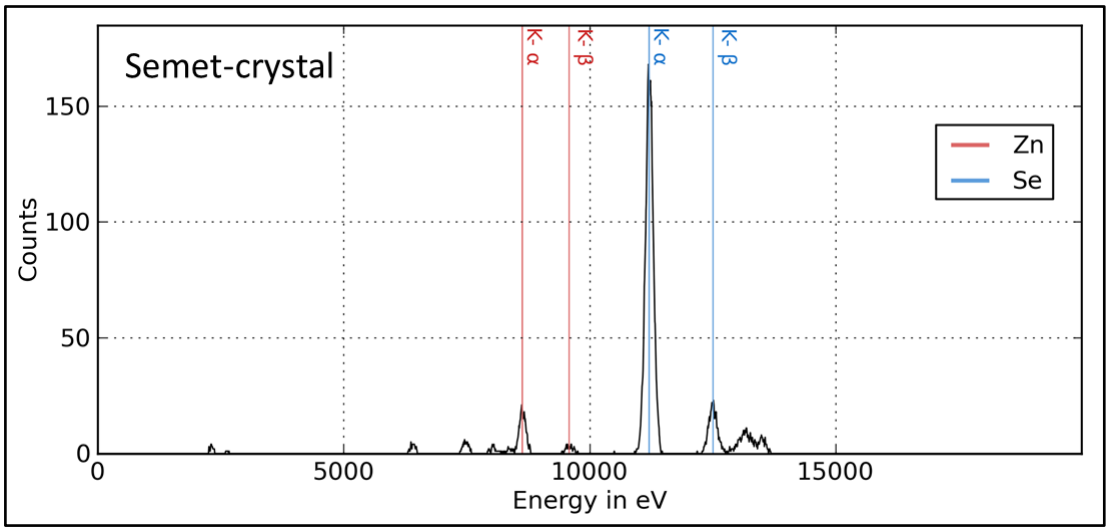


Figure 79: Fluorescence scan carried out with a selenomethionine derivative crystal at BESSY BL 14.1 (Mueller et al., 2012). Besides the selenium signal (blue line) a clear signal for zinc could be identified (red line).

Following this, a fluorescence scan at the zinc absorption edge was performed (figure 80). Two datasets collected around the zinc edge showed presence of anomalous signal in data collected at high energy remote and the absence of it below the edge (table 38). As the anomalous signal was not sufficient to solve the structure (data not shown), anomalous data were collected at the selenium edge (table 37).

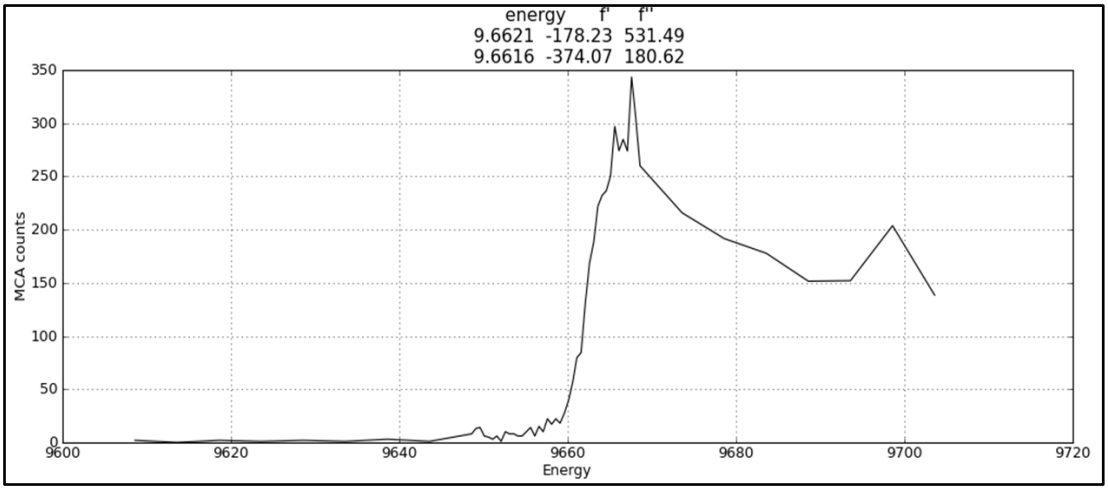


Figure 80: Fluorescence scan around the calculated zinc absorption edge carried out at BESSY II (Mueller et al., 2012).

Located heavy atom sites (Sheldrick, 2010) were used as additional input for AutoSol (Zwart et al., 2008; Adams et al., 2010) used to solve the structure. Manual model improvements were done using COOT (Emsley et al., 2010) alternating with automated refinement by phenix.refine (Adams et al., 2010) (table 39).

Table 37: Data collection statistics for selenomethionine derivative crystals

Data collection	SAD-PA5506[#*]	PA5506[#*$] peak	PA5506 infl.[#*$]
Detector	Pilatus 6M	Pilatus 6M	Pilatus 6M
Wavelength (Å)	0.979531	0.97981	0.979927
Resolution range (Å)[+]	42.3 – 1.87 (1.91 - 1.87)	48.3 – 2.2 (2.27 - 2.2)	46.0 – 3.0 (3.19 - 3.0)
Space group	$P4_12_12$	$P2_12_12_1$	$P2_12_12_1$
Unit cell parameters (Å)	70.9, 70.9, 157.3	66.9, 101.4, 109.9	66.9, 101.3, 110.0
Total no. of measured reflections[+]	432273 (30108)	1033355 (91236)	191495 (31325)
Unique reflections[+]	34127 (2268)	38830 (3311)	15511 (2451)
Multiplicity[+]	12.7 (13.3)	26.6 (27.6)	12.3 (12.8)
Anomalous multiplicity[+]	6.6 (6.8)	13.9 (14.4)	6.6 (6.7)
Mean I/σ(I)[+]	18.5 (2.3)	17.8 (4.0)	10 (3.5)
Anomalous completeness (%)	99.8 (100)	100 (100)	99.8 (98.7)
Completeness (%)[+]	100 (100)	100 (100)	99.8 (99.1)
Matthews' coefficient	3.17 monomer 1.58 dimer	2.99 dimer 1.99 trimer 1.49 tetramer	2.99 dimer 1.99 trimer 1.49 tetramer
Average Mosaicity[°]	0.13	0.07	0.17
R_{merge}[+] (%)	10.2 (141)	16.2 (126)	25.5 (116)
R_{meas}[‡] (%)	10.6 (152)	16.5 (126)	26.7 (121)
R_{pim}[$] (%)	3 (41.4)	3.2 (23.9)	7.7 (34.4)
CC(1/2)[+]	0.999 (0.732)	0.999 (0.923)	0.987 (0.776)

[#]Datasets were collected from single crystal
[$]peak and inflection data were collected from the same crystal
[*]Friedel mates were treated as separate reflections
[+]Values in parentheses refer to the highest resolution shell.
[°]Mosaicity reported by XDS, (Kabsch, 2010)
[‡]$R_{meas} = \Sigma_{hkl} (N/(N-1))1/2 \Sigma_i | I_i(hkl) - < I(hkl) > | / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where N is the number of observations of the reflection with index hkl and I_i is the intensity of its i^{th} observation.
[$]$R_{pim} = \Sigma_{hkl} (1/(N-1))1/2 \Sigma_i | I_i(hkl) - < I(hkl) > | / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where N is the number of observations of the reflection with index hkl (Weiss, 2001).

Table 38: Data collection statistics for data collected at around the zinc edge and for native data

Data collection	PA5506 Zn $_{hrem}$ $^{\#*\rho}$	PA5506 Zn $_{lrem}$ $^{\#*\rho}$	Native PA5506$^{\#}$
Detector	Pilatus 6M	Pilatus 6M	Pilatus 6M
Wavelength (Å)	1.282827	1.286153	0.918409
Resolution range (Å) $^{+}$	46.0 – 2.20 (2.27 – 2.20)	46.1 – 2.24 (2.31 – 2.24)	48.2-1.99 (2.04-1.99)
Space group	$P2_12_12_1$	$P2_12_12_1$	$P4_12_12$
Unit cell parameters (Å)	66.9, 101.5, 110	66.9, 101.5, 110	71.3, 71.3, 166.0
Total No. of measured reflections^{+}	426787 (34401)	259911 (21956)	217558 (15620)
Unique reflections^{+}	38723 (3299)	35850 (3242)	30325 (2107)
Multiplicity^{+}	11 (10.4)	7.1 (6.8)	7.2 (7.4)
Anomalous multiplicity	5.8 (5.4)	3.7 (3.5)	--
Mean I/σ(I) $^{+}$	11.5 (2.5)	10.7 (2.4)	20.1 (2.0)
Anomalous completeness (%)	99.9 (99.9)	99.4 (94.4)	--
Completeness (%)$^{+}$	99.9 (99.9)	99.5 (95.7)	100 (100)
Average Mosaicity $^{\circ}$	0.07	0.08	0.10
Matthews' coefficient	2.99 dimer 1.99 trimer 1.49 tetramer	2.99 dimer 1.99 trimer 1.49 tetramer	3.38 monomer 1.69 dimer
R_{merge} $^{+}$ (%)	14.7 (102)	10.7 (72.7)	6.7 (103.2)
R_{meas} ‡ (%)	16.1 (113)	12.4 (85.5)	7.2 (110.9)
R_{pim} $^{\$}$ (%)	6.5 (34.9)	5.6 (32.6)	2.7 (40.4)
CC(1/2)$^{+}$	0.997 (0.823)	0.997 (0.813)	0.999 (0.692)

$^{\#}$Datasets were collected from single crystal

$^{\rho}$peak and inflection data were collected from the same crystal, datasets were left unrefined.

*Friedel mates were treated as separate reflections

$^{+}$Values in parentheses refer to the highest resolution shell.

$^{\circ}$Mosaicity reported by XDS, (Kabsch, 2010)

$^{\ddagger}R_{meas} = \Sigma_{hkl} (N/(N-1))1/2 \Sigma_i | I_i(hkl) - < I(hkl) > | / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where N is the number of observations of the reflection with index hkl and I_i is the intensity of its i^{th} observation.

$^{\$}R_{pim} = \Sigma_{hkl} (1/(N-1))1/2 \Sigma_i | I_i(hkl) - < I(hkl) > | / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where N is the number of observations of the reflection with index hkl (Weiss, 2001).

Table 39: Refinement statistics

Refinement statistics	PA5506_SeMet_SG92	PA5506_SeMet_SG19	PA5506 native
Resolution (Å) [+]	42.3 – 1.87 (1.94 – 1.87)	48.3 – 2.20 (2.27 – 2.20)	43.7 – 1.99 (2.06 – 1.99)
Space group	$P4_12_12$	$P2_12_12_1$	$P4_12_12$
$R_{work\ (\%)}$ [+]	17.2 (25.6)	19.6 (26.6)	20.0 (28.3)
$R_{free\ (\%)}$ [+]	21.1 (28.5)	23.4 (27.1)	22.9 (29.8)
No. of non-hydrogen atoms	3132	5900	2895
Protein	2944	5751	2823
Water	186	87	72
# chains / ASU	2	4	2
Matthews coefficient	2.31 ≙ 46 % solvent	2.18 ≙ 43 % solvent	2 ≙ 46% solvent
Average B-factors (A^2)	36.7	47.1	44.2
Protein	36.6	47.0	44.4
Water	37.6	37	37.7
R.m.s. deviations			
Bond lengths (Å)	0.009	0.004	0.003
Bond angles (°)	1.07	0.79	0.73
Ramachandran plot [!]			
Favored regions (%)	98	98	97.5
Outlier (%)	0	0	0.0
Validation score[&]			
Clashscore	2.4	2.1	1.08
MolProbity score	1.02	1.42	1.04

[+]Values in parentheses refer to the highest resolution shell.
[&]As reported by MolProbity (http://molprobity.biochem.duke.edu/ (Chen et al., 2010)

[!] Percentage of residues in the Ramachandran favoured region was calculated using MolProbity.

5.4.4. PA5506 crystallized in truncated form

The *Pseudomonas aeruginosa* PA5506 consists of 285 amino acids. Past data collection the Matthews's coefficient (Matthews, 1968) was calculated for every dataset. Interestingly, no plausible coefficient could be calculated for any of the datasets using the molecular weight of PA5506, calculated to 32.1 kDa. AutoSol (Zwart et al., 2008) built two and four protein chains to the asymmetric unit to the different space groups, respectively (table 4). Structural inspection after initial refinement revealed the reason for incorrect assignment of the Matthews coefficient. Despite the fact that full-length protein was used for crystallization (confirmed by mass spectrometry) no electron density could be observed for the first ~95 amino acids. Assumption that the whole N-terminal domain is probably unstructured and therefore not visible in the electron density was proven false by analysis of crystal packing. Molecule arrangement in both space groups does not allow the placement of an extra domain of more than 90 amino acids. Initially, it was thought that the protein comprises a larger unstructured region, however, which was indicted as false by predictions of high disorder tendency (figure 81) (Ishida & Kinoshita, 2007).

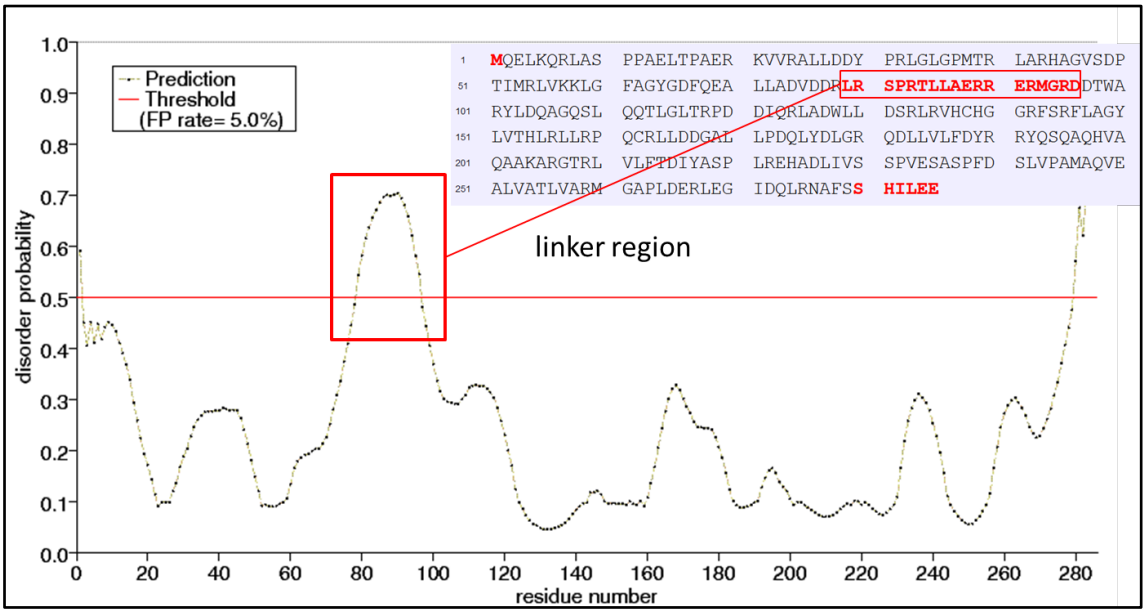


Figure 81: Disordered protein region identified by the Protein DisOrder prediction System (PrDOS) (Ishida & Kinoshita, 2007)

A disorder was only predicted for a region comprising amino acid 79 through 96 (Ishida & Kinoshita, 2008) which is probably a linker region connecting the two protein.

The protein content of isolated and washed crystals analyzed by SDS-PAGE revealed a lower molecular weight of around 20 kDa (figure 82B). To identify the disposition of the cleaved protein fragment, collected 'crystal supernatant' was applied to mass spectrometry. Results indicate the presence of smaller protein fragments in the solution at around 9 kDa (figure 82A). However, a clear band that would reflect a distinct fragment is not observed on a gel. Furthermore, the intensity of this peak is rather low compared to ones at an m/z of ~4000 that are attributed to PEG molecules, contained in the mother-liquor composition.

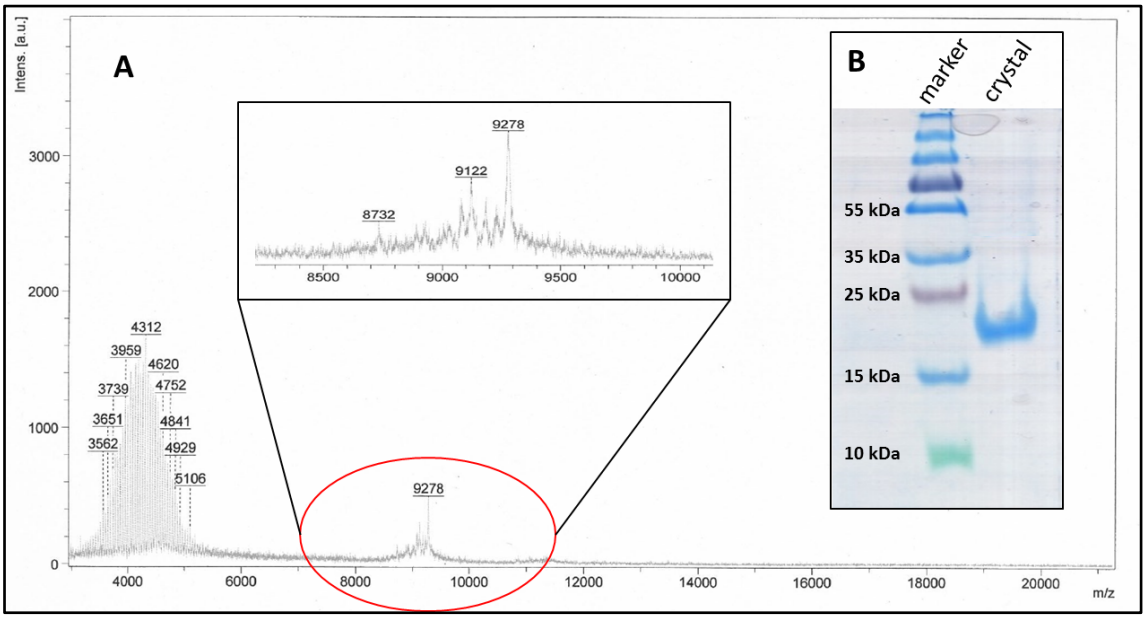


Figure 82: Analysis of crystal content (B) and mass analysis of 'crystal supernatant' (A).

Therefore, the protein is either degraded over time or, however, very unlikely in this case the cleavage is induced by components of the mother liquor used for crystallization. Hence, a comparative gel analysis using samples of both, pure native and SeMet labeled protein, stored for four weeks at 4°C prior to analysis, revealed a degradation of full length protein without any further external impact (figure 83), suggesting that the protein is cleaved before it crystallized and the N-terminal domain is degraded.

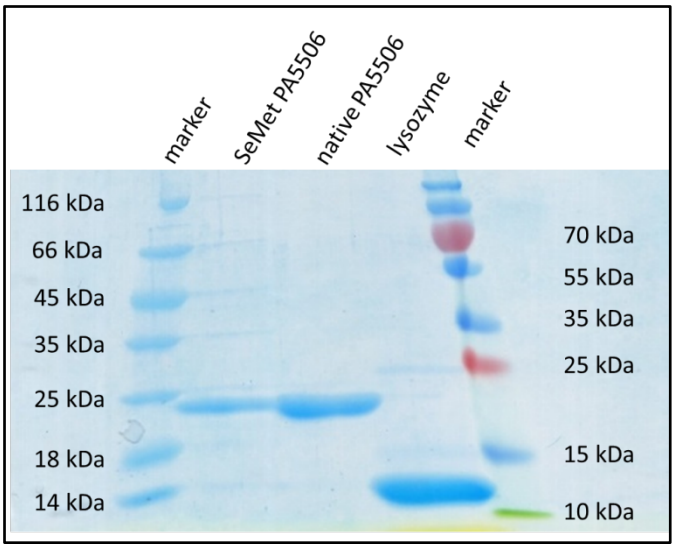


Figure 83: Analysis of PA5506 after four weeks storage in the fridge. Lysozyme was applied as a control.

5.4.5. Structure of PA5506

The structure of the selenomethionine (SeMet) labeled protein was solved in space group $P2_12_12_1$ at 2.20 Å resolution and refined to R_{work} = 19.6% and R_{free} = 23.4% as well as in space group $P4_12_12$ at 1.87Å resolution with R-values of 17.2 % and 21.1 % for R_{work} and R_{free}, respectively. The asymmetric unit of the $P4_12_12$ crystals contained two protein chains and four subunits were present in the asymmetric unit of the $P2_12_12_1$ crystal (figure 84).

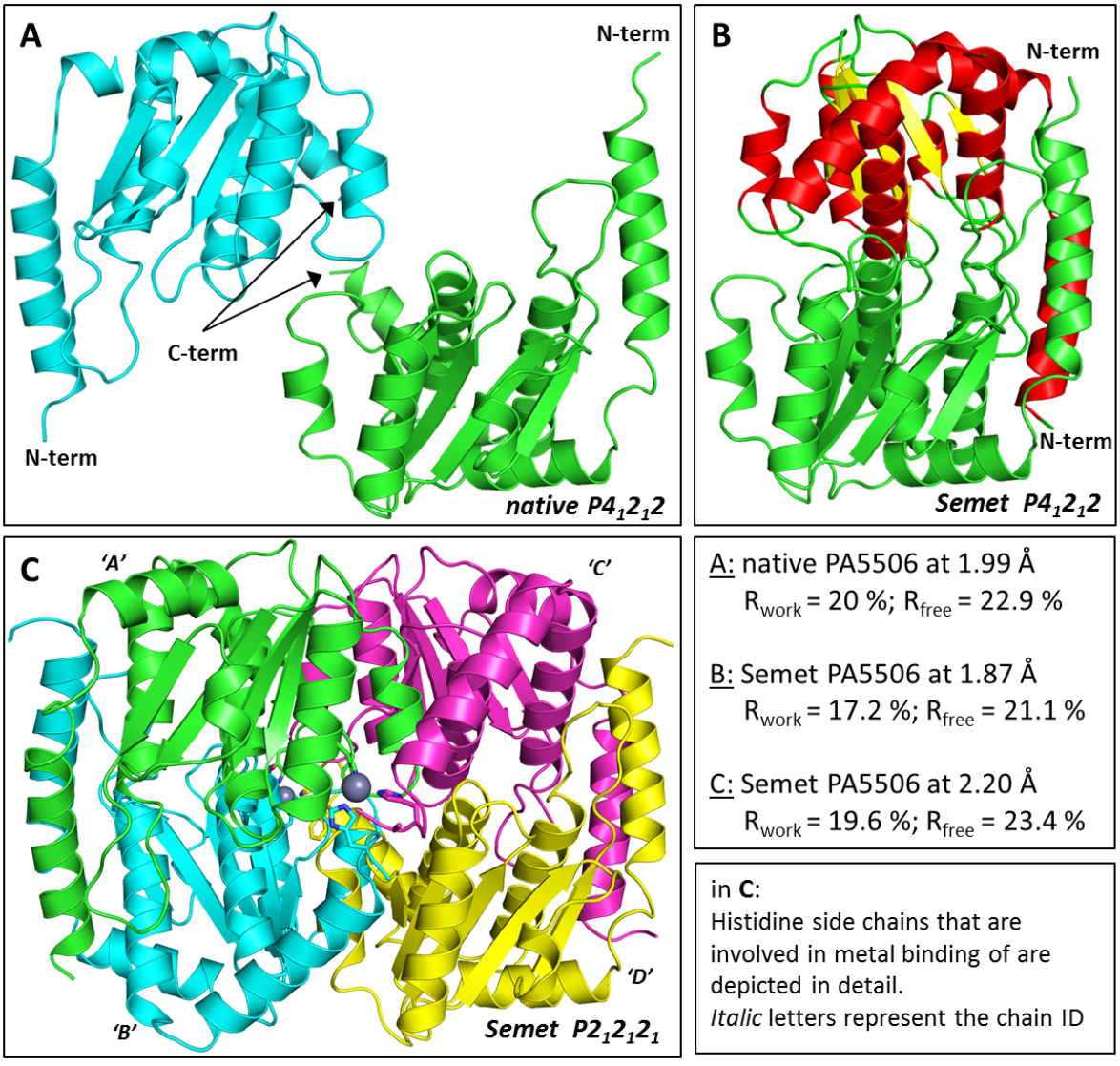


Figure 84: Content of the asymmetric unit of all solved structures. The N- and C-termini are labeled accordingly. Both native (A) and SeMet crystal (B) belong to space group 92 and contain 2 chains, the secondary structure elements are coloured red and yellow (B). The structure solved in space group 19 (C) did show four monomers in the asymmetric unit.

All residues of the protein structure fall into the allowed and 98% in the favored region of the Ramachandran plot with a MolProbity (Davis et al., 2004), used for structural validation reported score values of around 1.2 (Chen et al., 2010).

The overall fold of the protein consists of a parallel 5-fold beta blade which is clamped by alpha helices on either side with two additional long helices, one at the N- and the C-terminus, stretching of the core fold, α_1 and α_8. This fold is also found in flavodoxin-like α-β-α three layer sandwich fold proteins (figure 84 and 85).

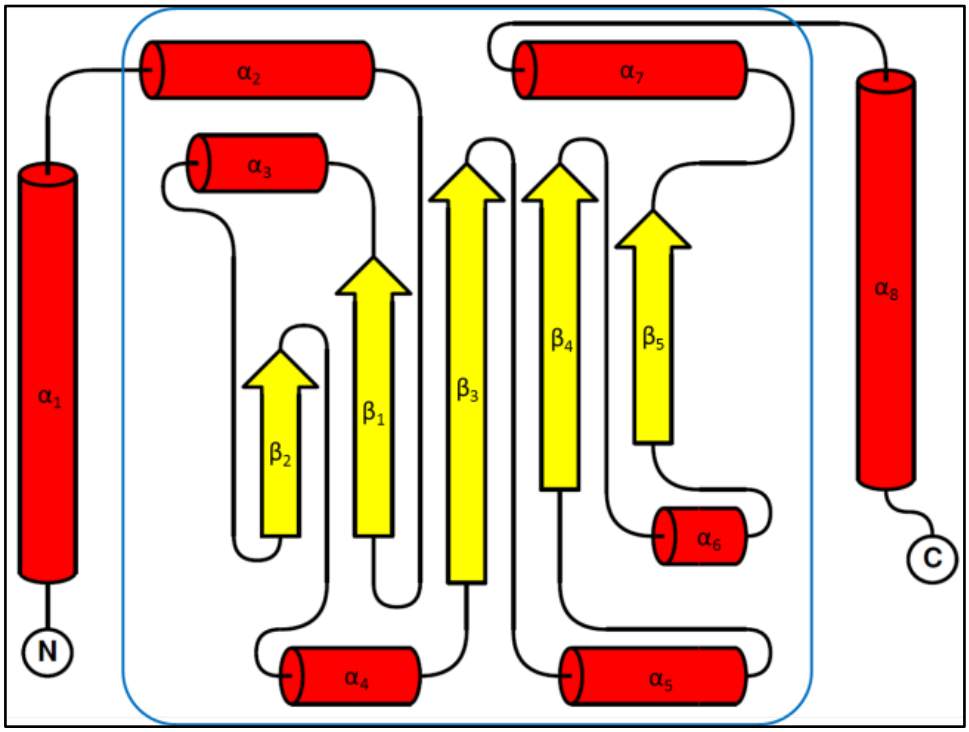


Figure 85: Overall topology of a PA5506 monomer shows a three layer α-β-α-fold. The core fold is indicated by the blue box with N- and C-terminal extending long helices.

The two monomers of PA5506 crystallized in space group $P4_12_12$ are complemented by two symmetrical molecules to the tetrameric protein structure. The protein assembly comprises a dimer of dimers (chain A and B form dimer 1 and chain C and D form dimer 2) related by a two-fold axis resulting in a tetramer with 222 symmetry. Each monomer has a surface exposed area on average of about 10500 A^2. The protein dimers comprising chain A/B and C/D are bound to each other mainly by a hydrogen bond network and salt bridges as well as few hydrophobic interactions between α_1, α_3 and α_7 of one monomer

and their counterparts from the second monomer with a total buried protein surface of 4500 Å². The interactions between the two dimers AB and CD are mostly hydrophobic with only a few hydrogen bonds. The buried area between A/B and C/D contributes about twice to the total buried area of the tetramer. The four subunits of the tetramer within the protein structures are very similar to each other, with r.m.s.d.s of 0.24 Å for both structures in $P4_12_12$ and 0.22 Å for the structure in $P2_12_12_1$. Protein dimers are assembles in a head-to-head manner with major contributions of the extended N-terminal helix of either monomer. The arm like α_1 composition inserts into a groove, generated by α_1', a loop region connecting β_5' to α_7' and α_7' itself (figure 86).

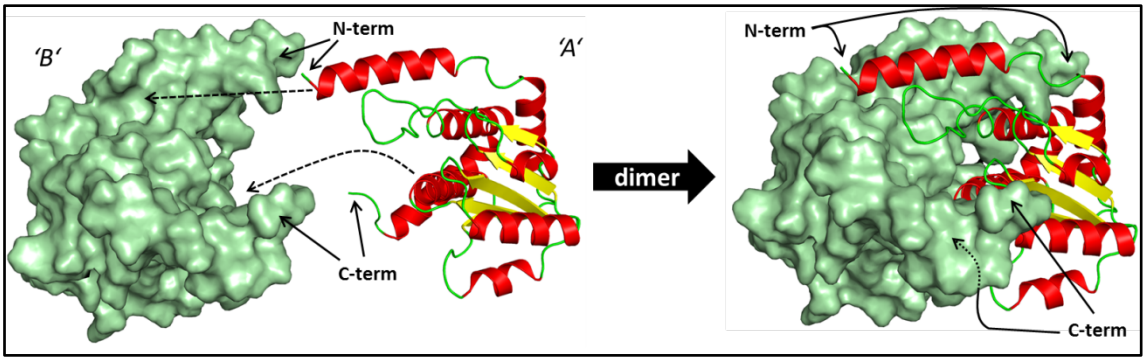


Figure 86: Dimerization of two PA5506 subunits in a 180°-turned organization

Extensive hydrogen bond networks as well as 11 salt bridges stabilize the overall structural conformation (table 40).

Table 40: Analysis of engaged interfaces

##	Interfacing structures	Buried surface [Å²]	Δ^iG [kcal/mol]	# HB	# SB
1	A + B	2699	-45.9	13	11
2	C + D	2697	-45.8	6	9
3	A + D	1198	-11.2	4	8
4	B + C	1159	-11.4	5	10
5	ABCD	26612	-163	-	-

Two of the dimers arrange in a 180° turned manner where the small helix α_4 of one monomer inserts into a pocket generated by α_6' the C-terminal α_8' of the other monomer in a 'key-lock' manner (figure 87).

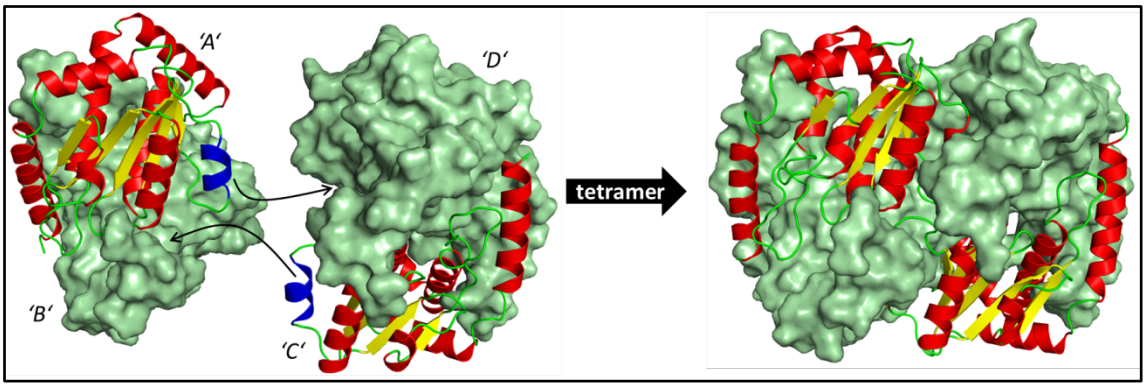


Figure 87: Tetramer assembly – a dimer of dimers. The interlocking 'key' is colored in blue.

5.4.6. PA5506 has flexible termini

The overall architecture of all protein structures is very similar except slight differences on the N- and C-terminus. For clarity the altering termini are depicted on a single protein chain of each structure (figure 88).

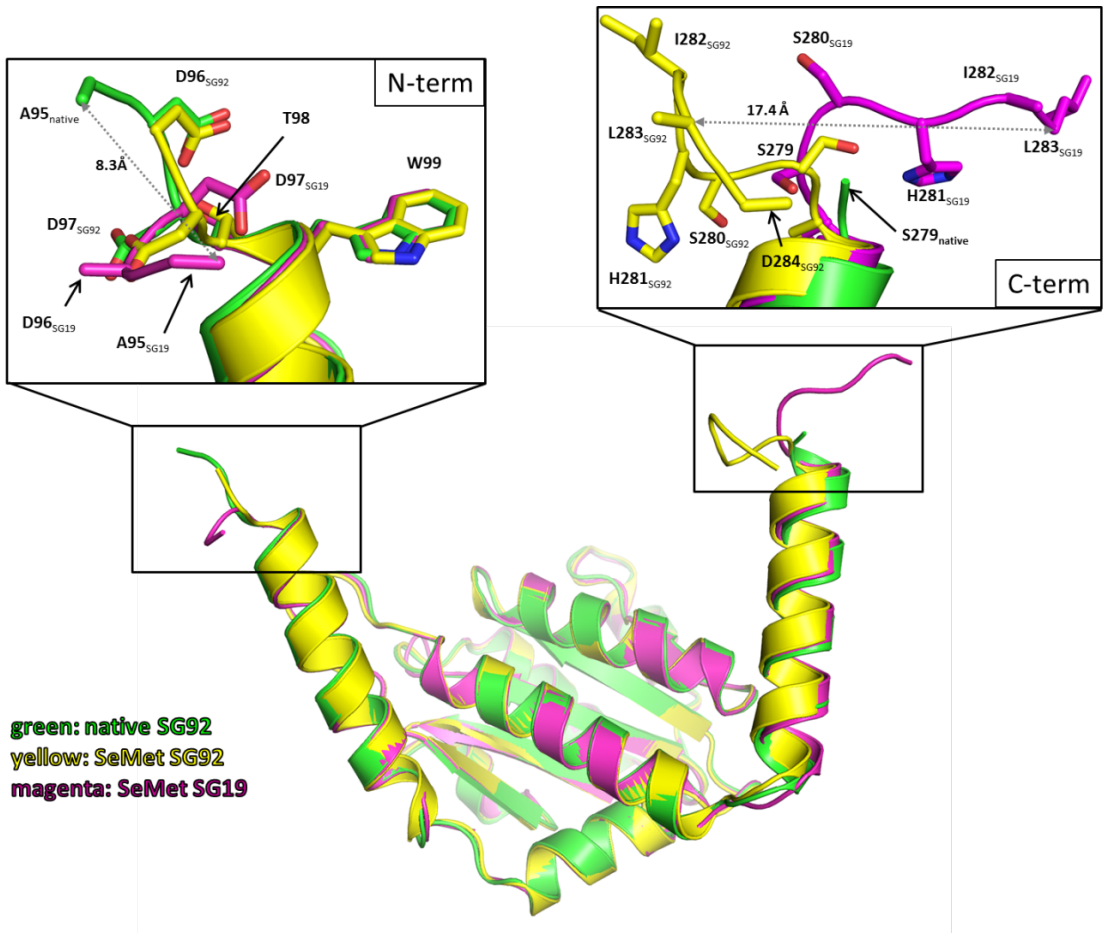


Figure 88: Superposition of single protein chains. The core structure is almost identical, the differences are found on N- and C-terminus region. Significant changes are depicted in the magnification of the termini. Amino acids are labeled according to the nomenclature of the full length protein, starting with amino acid 1 through 285.

As PA5506 contains a predicted N-terminal DNA-binding (DBD) domain, which lacks in all crystal structures, conformational alterations observed at the N-terminus might not be of significant importance and be caused by crystal packing artefacts. Furthermore, this terminus will in the full-length protein never be exposed to the surrounding environment.

Alterations on the C-terminus are more relevant as terminal residues, often not considered essential, have been shown in some proteins to achieve a major role in regulation of the enzyme. For example, the C-terminus of ERK5 a unique mitogen-activated protein (MAPKs) was shown to be involved in auto-inhibition of the protein's activity (Buschbeck & Ullrich, 2005). Also, occasional a conformational change in the C-terminal region controls the activity of a huge enzyme complex (de la Mata & Kornblihtt, 2006), as small part of the C-terminal domain of the RNAII polymerase was reported to be essential for its stability (Chapman et al., 2004). The C-terminus shows major differences amongst all three structures. The homogenous chain branches of into three different directions at phenylalanine 278. Native- and $SeMet_{SG19}$-main chains are diverging generally in the same, whereas the $SeMet_{SG92}$ chains splits in a different direction (figure 88).

5.4.7. Metal binding involves a C-terminal rearrangement

The two structures of PA5506 clearly indicate that the protein is a metal binding protein, which was proven by a fluorescence scan and the observation of anomalous signal in data collected around the zinc edge.

In contrast to a homologous SIS-member, the sedoheptulose-7-phosphate isomerase GmhA from *Burkholderia pseudomallei* which binds the metal in the 'heart of the active site' where it participates in reaction mechanism (Harmer, 2010), metal binding in PA5506 is realized on the surface of the tetramer.

The metal free PA5506 crystallized in space group $P4_12_12$ (figure 89) and the metal bound protein structure was determined in space group $P2_12_12_1$ (figure 90).

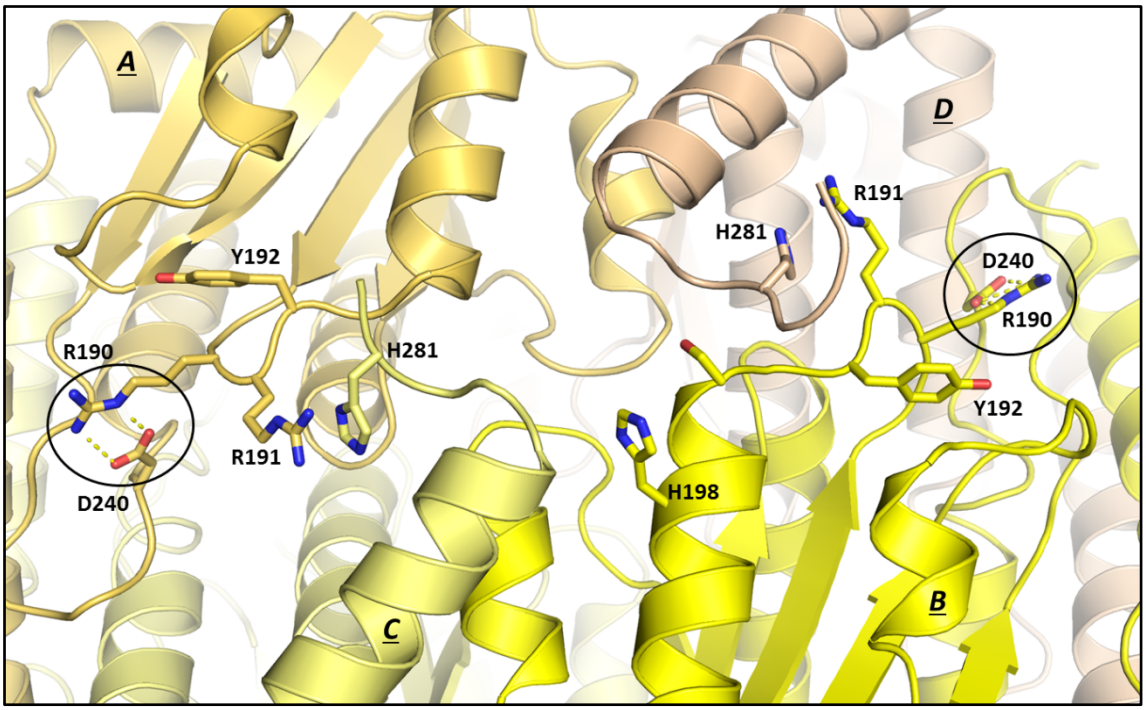


Figure 89: Conformation of metal binding site in absence of zinc. The structure in space group $P4_12_12$ with the four chains coloured individually and labeled in underlined italic letters.

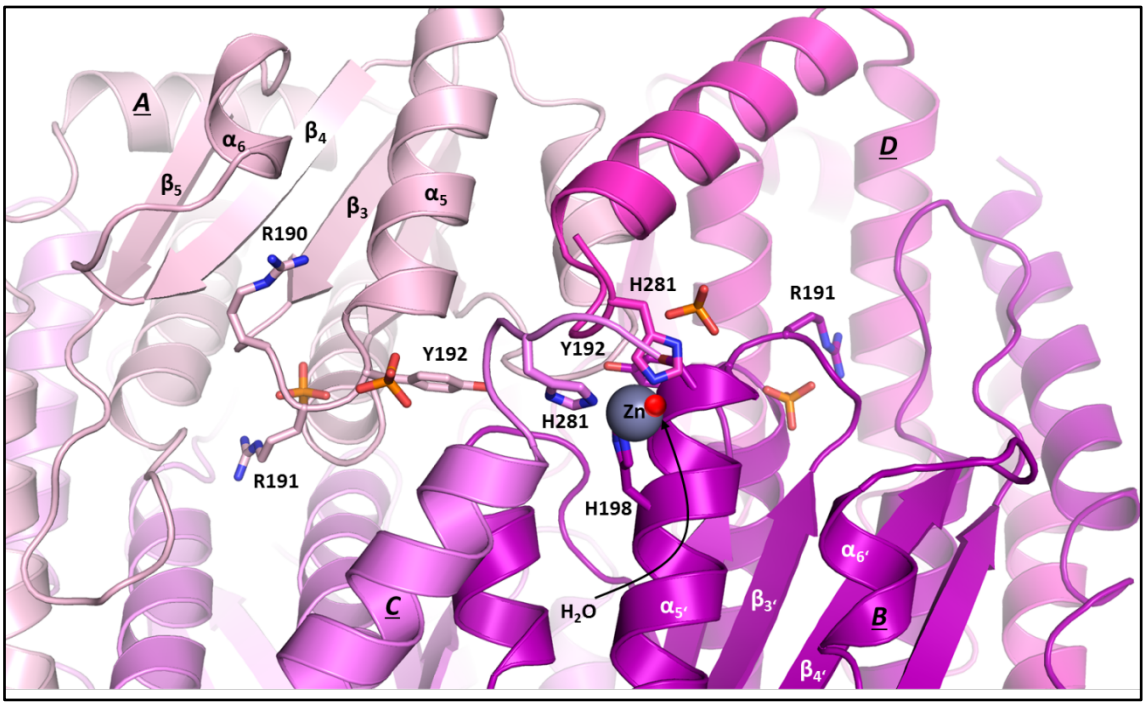


Figure 90: Zinc binding site in the $P2_12_12_1$ crystal form. Chains are indicated by underlined italic letters and relevant residues are individually labeled.

Both structures are very similar in the overall fold and superpose with an r.m.s.d of 0.52 Å. However, structural changes were identified in all four subunits happens upon metal binding, involving the C-terminal tail as well as a loop region joining β_3 and α_5. Histidine residues coordinate the metal ion in a tetrahedral manner (figure 91).

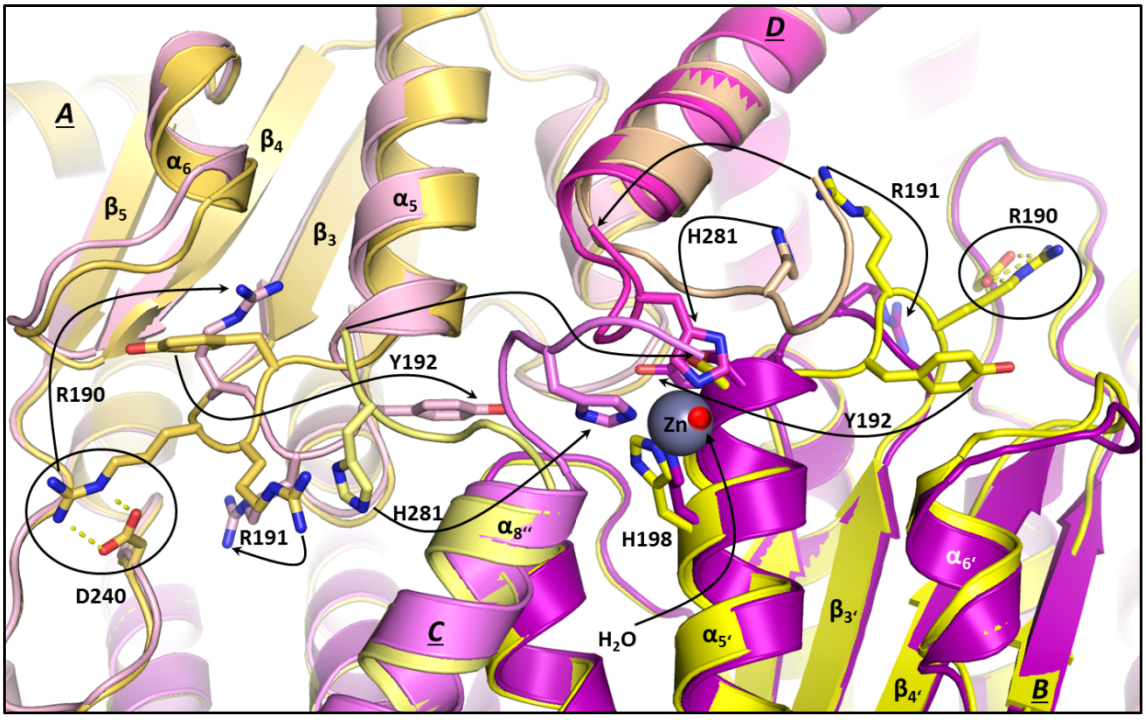


Figure 91: Superposition of metal-free (yellow) and metal bound structure (magenta). Major rearrangements are noted upon metal binding. The phosphate molecules (figure 17) were removed for clarity. Chains IDs are indicated by italic bold letters.

The metal-free structure is named the 'open' conformation and the metal bound is termed 'closed' conformation. Two major differences between the distinct structures are apparent. First, the loop region connecting β_3 to α_5, the 'Tyr/Arg-switch loop', undergoes a structural inversion. In the open state, an intramolecular salt bridge between R190 and D240 stabilizes the loop position. Upon metal binding this contact is lost (figure 18, circle). Moreover, tyrosine 192 flips about 180° into the center of the tetramer ~18 Å from its original position, losing hydrogen bond contacts to D215 and the backbone of S219. R190 flips in the close conformation into the position previously occupied by Y192. Secondly, the overall positioning of the C-terminus is different (chain C and D, figure 18). The conformation in the open state is stabilized by intramolecular hydrogen bonds between the side chain of S280 and main chain of R275 as well as water mediated contacts

between the side chains of H281 and R275. Histidine 281 (chain C), which is involved in metal coordination, moves its positon about 16 Å to complete the metal binding site. Finally, the zinc atom is coordinated by histidine 281 of chain *C* and *D* and histidine 198 of chain *B* on one side of the tetramer. The second metal is coordinated in the exact same manner by the C-terminal residues of chain *A* and *B*. The third histidine originated from chain *D*. The tetragonal coordination is completed by a water molecule located in a distance of 2.6 Å (figure 92B).

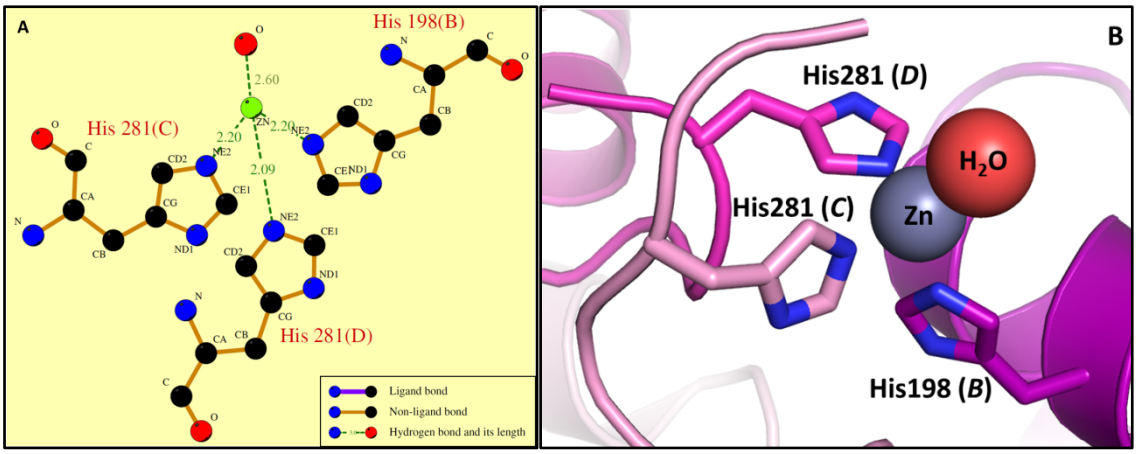


Figure 92: Metal binding site of PA5506. The zinc atom is found in tetrahedral coordination (B), realized by the side chains of three histidine residues and a water molecule (A,B) (Laskowski & Swindells, 2011).

5.4.8. Ligand binding site identification

Observed phosphate molecules, though originating from the crystallization buffer of the $P2_12_12_1$ crystal, are of relevance for the structural changes. Not observed in the apo structure, one phosphate molecule stabilizes the conformation of arginine 191 by generating a hydrogen bond network within the loop region. Structural comparison with a homologous structure from *Bacteroides fragilis* (*PDB* ID: 3ETN) (Chiu et al., 2014) and *Pseudomonas aeruginosa* (PDB ID: 1X92) (Taylor et al., 2008), co-crystallized with its phosphorylated substrate, indicated this phosphate as putative conserved molecule. In fact, this phosphate is located close to predicted active site residue serine 144 and occupies almost the identical position (figure 21). The ligand-binding site was located by structural comparison to structures of related proteins. Located in a cleft at the tetramer

interface it is composed of residues from α_3 and the loop region from β_3-α_5 originating from chain *A* and the C-terminal helix α_8 from the neighboring chain *B*.

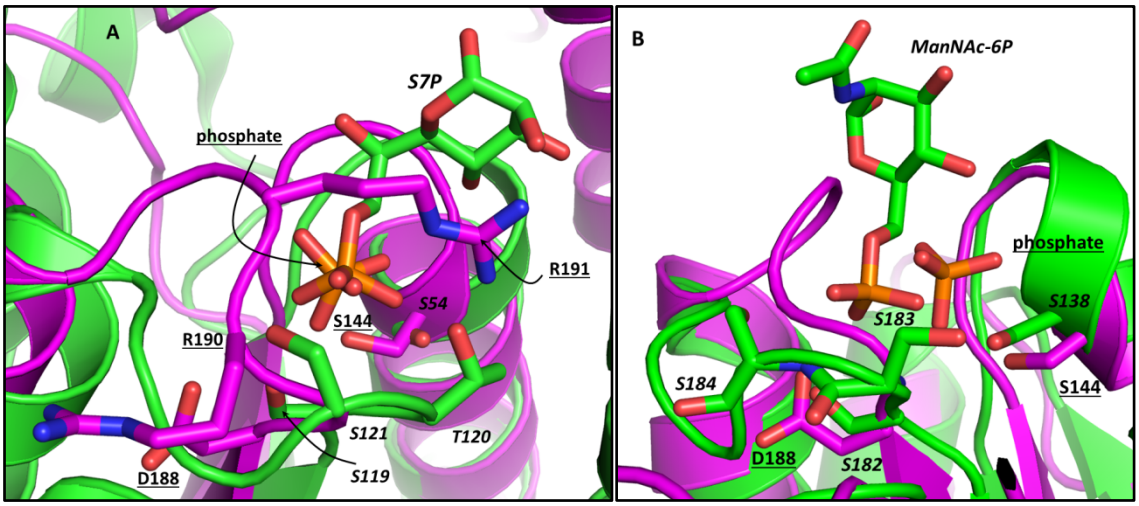


Figure 93: Superposition of the putative ligand-binding site of PA5506 with representative homologous structures. A sugar isomerase (A) from *Pseudomonas aeruginosa* (green) (*PDB*: 1X92) superposed with PA5506 (magenta) (A) and a transcription regulator from *Vibrio vulnificus* (green) (*PDB*: 4IVN) (B). Residues of PA5506 are labeled with underlined numbers and those of the homologous structures are presented in italic format, S7P = Sedoheptulose-7-phosphate and ManNAc-6P = 2-(Acetyl-amino)-2-desoxy-6-O-phosphono-alpha-D-mannopyranose.

The phosphate in homologous structures is coordinated by a loop connecting β_3 to α_6, previously described as the 'Ser/Thr-rich loop' (Chiu et al., 2014). This loop is conserved amongst arabinose-phosphate isomerases, however, is replaced by the 'Tyr/Arg-switch loop' in PA5506. Structural reorganization of this loop is crucial for completion of the active site (figure 93A) and phosphate binding by the amide backbone of S144, F143, R142, R191 as well as the side chain of R191 in its 'close' conformation (figure 94).

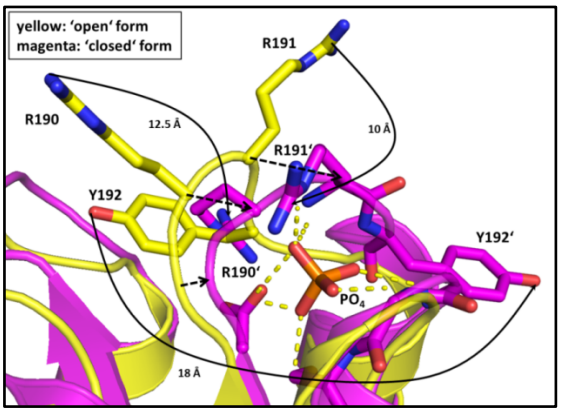


Figure 94: Superposition of both active site conformations. The phosphate ion is bound to the $P2_12_12_1$ structure (magenta). Hydrogen bonds are indicated by the dotted yellow lines. The open conformation is shown in yellow. Residues labeled with prime represent the corresponding residue within the other state.

5.4.9. Structure in context

While this study was carried out, PA5506 was identified as an RpiR-family transcription regulators homolog, named *qapR* (Tipton et al., 2013). It regulates the expression of the *qap*-operon comprising three downstream genes *PA5507*, *PA5508* and *PA5509* in a negative way (Tipton et al., 2013). Deletion of *PA5506* and therefore expression of the *qap*-operon results in decreased level of PQS which subsequently decreased the levels of pyocyanin, as PQS signaling is its primary regulator (Blankenfeldt & Parsons, 2014; Welsh et al., 2015). The PQS molecule as a positive feedback inducer influence on its own synthesis, thus the transcription level of the biosynthetic operon is found to be much lower and could be restored in the *qap* mutant by the addition of PQS (Tipton et al., 2013). In the Δ*qapR* mutant strain qapR is truncated down to a core fragment, lacking the first 30 amino acids of the HTH-motive and the last ten C-terminal amino acids. Obviously, binding to DNA will be affected by the degenerated DNA-binding motive and hence more interesting is the C-terminal truncation, as structural results suggest that the C-terminus is involved in metal coordination and therewith might be involved in regulation of protein's activity.

A BLAST search identified a high conservation of PA5506 (QapR) amongst *Pseudomonas* species but also a few orthologs in other species (Altschul et al., 1997). Amongst the identified RpiR-members only one is structurally characterized in full length, the dimeric structure NanR, whose C-terminal domain matches QapR with a deviation of 3.5 Å (*PDB*: 4IVN) (Hwang et al., 2013) as identified by HMMER (Finn et al., 2011).

Metal binding was only reported for one SIS-domain containing protein which, however, follows a metal catalyzed reaction mechanism (Harmer, 2010). Interestingly, the composition of the C-terminus of PA5506, which is involved in zinc binding was identified as unique by alignment of 1000 sequences. As transcription regulator, QapR probably does not act as sugar isomerases but rather binds a certain type of ligand molecule, which upon binding alleviates the repressive effect of QapR. Based on the structural findings and comparison to existing homologs, phospho-sugar molecules or their derivatives are postulated as putative ligands.

In close proximity to the *qap*-operon, a predicted membrane bound transporter is coded, PA5510. Tipton and coworker did not find this gene to be expressed simultaneously with

the *qap*-operon, however, available transcriptome data of more than 100 *Pseudomonas* isolates provide strong evidence of co-expression along with members of the *qap*-operon.

5.5. Structural analysis of PA5507

5.5.1. PA5507 sequence analysis

The protein sequence was retrieved from the *Pseudomonas* Genome Database (PGD) entry PA5507. The hypothetical cytosolic protein has 31 putative orthologs among *Pseudomonas species* (Winsor et al., 2011) and is embedded in the *qap*-operon, covering the genes *PA5506-PA5509* (Tipton et al., 2013). Structural prediction by PHYRE (Kelley & Sternberg, 2009) identified PA5507 as member of the α/β-hydrolase super family with strong homology to the cysteine hydrolase fold family (figure 95).

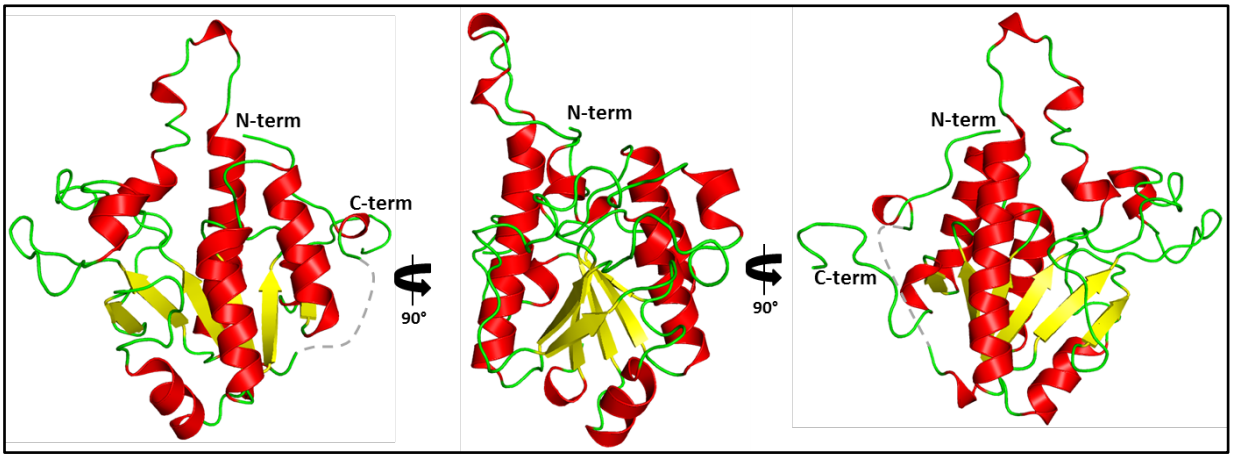


Figure 95: Structural prediction result by PHYRE using PA5507 sequence as query. Beta strands a colored in yellow while alpha helical part a stained in red. The C- and N-termini of the predicted protein chain are indicated with C-term and N-term, respectively. The grey dashed line connects a part which could not be predicted by the automated online server PHYRE (Kelley & Sternberg, 2009).

This superfamily covers five subclasses of which the N-carbamoylsarcosine amidohydrolase (CSHase) proteins are involved in creatinine catabolism (Wyss & Kaddurah-Daouk, 2000) and members of the nicotinamidase family converts nicotinamide into nicotinic acid and ammonia (Lemaitre et al., 2001; Smith et al., 2011). The third family member are the isochorismatases that catalyze the conversion of isochorismate to 2,3-dihydroxy-2,3-dihydrobenzoate and pyruvate via hydrolysis of a vinyl ether bond (Parsons et al., 2003). Last, the nicotinamidase-related proteins, possessing the same catalytic triad than nicotinamidases hence, do not convert nicotinamide. Finally, the family of YcaC-like proteins that shares a high homology to the hydrolase YcaC from *Escherichia coli* (Colovos et al., 1998). Sequence comparison did not reveal a distinct family class that PA5507 might be member of, it only excluded the isochorismatase class because of the predicted active site cysteine residue. Common amongst all members is a

conserved cis-peptide bond in the C-terminal part of the protein. This conformation allows the generation of an oxyanion hole to stabilize the intermediate state of the substrate conversion during catalysis as it is part of a predicted catalytic triad: D25, K123 and the C156 (Marchler-Bauer et al., 2013). Concluding here, PA5507 is a member of the cysteine hydrolase family and might be involved in secondary metabolite biosynthesis, transport or catabolism, suggested by COGs annotation (Cluster of Orthologous Groups) (Tatusov et al., 1997; Wolf et al., 2012).

5.5.2. Structure solution and model building

Crystals of PA5507 were obtained in orthorhombic space group $P2_12_12$ (figure 96) with two monomers in the asymmetric unit. Summarized data collection statistics are reported in table 41.

Table 41: Data collection statistics for PA5507

Data collection	PA5507	PA5507 high res [ß]
Detector	Mar345 image plate	Pilatus 6M
Wavelength (Å)	1.5418	0.9724
Resolution range (Å)	20 – 1.66 (1.70 – 1.66)	42.13 – 1.36 (1.38 – 1.36)
Space group	$P2_12_12$	$P2_12_12$
Unit cell parameters (Å)	82.17 93.55 52.83	83.45 84.27 52.71
Total No. of measured reflections	172719 (6514)	559039 (25882)
Unique reflections	46851 (2122)	80401 (3955)
Multiplicity	3.7 (3.1)	7.0 (6.5)
Mean I/σ(I)	17.35 (2.58)	13.7 (2.3)
Completeness (%)	96.9 (82.3)	99.8 (99.8)
Average Mosaicity [°]	0.467	0.285
R_{merge} (%)	5.1 (47.2)	7.0 (71.9)
R_{meas} (%)	6.9 (57.6)	8.2 (85.0)
R_{pim} (%)	3.5 (31.3)	4.2 (44.6)
CC(1/2) [°]	0.998 (0.71)	0.999 (0.721)

[#]Data were collected from single crystal; [+]Values in parentheses refer to the highest resolution shell. [ß]Data collected at ESRF, ID23.1 (Nurizzo et al., 2006b), [°]Mosaicity and CC(1/2) reported by XDS (Kabsch, 2010; Karplus & Diederichs, 2012), [$]Rpim = Σhkl (1/(N – 1))1/2 Σi | Ii (hkl) – < I (hkl) > | / Σhkl Σi Ii (hkl), (Weiss, 2001).

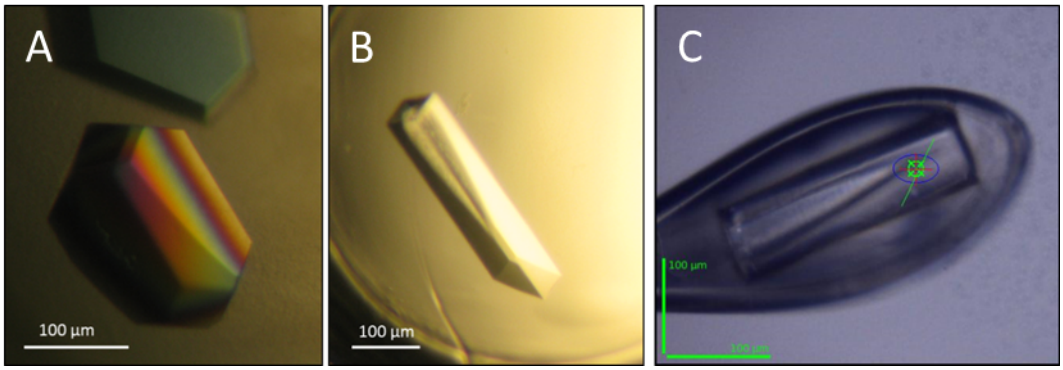


Figure 96: Shape of moderate (A) and high diffracting (B) crystals. Cryogenic mounted PA5507 crystal at ID23.1, ESRF (C).

The phase problem was solved by molecular replacement (MR) using the automated pipeline BALBES (Long et al., 2008). Structure solution revealed two amino acid chains in the asymmetric unit forming a homodimer. The Matthews coefficient was calculated to 1.86 $Å^3$ Da^{-1} with a solvent content of 34%. Manual model adjustments were done by COOT (Emsley et al., 2010), altering automated refinement using phenix.refine (Afonine et al., 2012). The final model was refined to 1.36 Å with a R_{work} of 14.9 % and an R_{free} of 17.1 %. MolProbity (Davis et al., 2004), reported a score of 0.61 with all protein residues in the allowed and 98.4% of in favored region of the Ramachandran plot (figure 97). Refinement statistics are summarized in table 42.

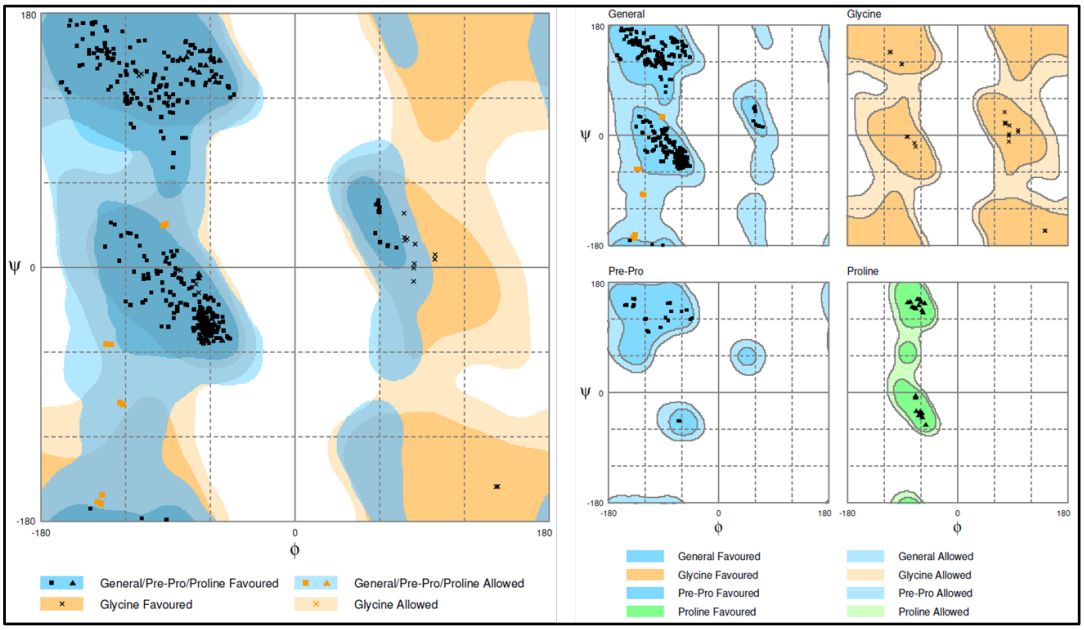


Figure 97: Ramachandran plot analyzing the structure of PA5507, generated by RAMPAGE (Lovell et al., 2003).

Table 42: Refinement statistics of PA5507

Refinement	PA5507 high res
Resolution (Å)	42.13 – 1.36 (1.38 – 1.36)
Number of reflections	80342 (7918)
R_{work} (%)	15.2 (24.7)
R_{free} (%)	17.1 (27.5)
# non hydrogen atoms	3744
Protein *	3355
Water	380
Average B-factors (Å²)	
Protein	17.0
Water	24.3
R.m.s. deviations	
Bond length (Å)	0.009
Bond angle (°)	1.23
Rotamer outlier (%)	0
Ramachandran plot (%)	
Favored region	98.4
Outlier	0
Validation score	
Clashscore	0.3
MolProbity[&]	0.61

* indicates the number of non-hydrogen, non-solvent atoms; [&]as reported by MolProbity ((Chen et al., 2010), http://molprobity.biochem.duke.edu/).

5.5.3. Structure of PA5507

The asymmetric unit contained a protein dimer (figure 98). The protein's architecture is similar to the 3-layer 'alpha-beta-alpha sandwich family', with the characteristic organization of 6 parallel twisted β-sheets clamped with flanking alpha helical bundles.

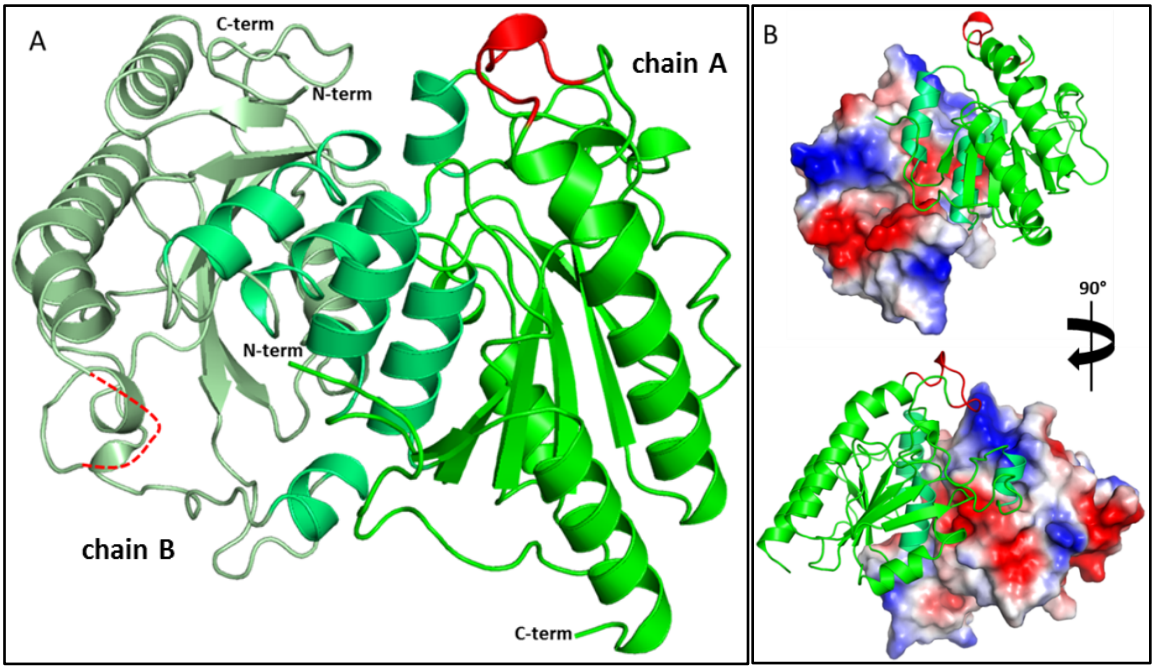


Figure 98: Two chains occupy the asymmetric unit of the crystal. Chain B is stained in pale green and chain A in green with the red colored loop region (A). The alpha helices of each monomer contributing to the dimer interface are stained in lime green (7A). In 7B the dimer itself is mainly established by polar and charged contacts visible in 7B. The electrostatic calculations revealed a charged interface (7B).

The dimer interface between the two chains is realized by parallel cross-spanning alpha helices of each subunit creating a mostly polar protein dimer interface. In fact, analysis of crystal packing revealed a protein tetramer, which is assembled by chains from two neighboring asymmetric units. This is in congruence with results from size exclusion chromatography indicating a higher molecular weight.

PISA analysis (Krissinel & Henrick, 2007) predicted a solvent-accessible surface area of 9970 Å² for a monomer which is reduced by about ~2100 Å² upon tetramer assembly. Collected SAXS data clearly identified the protein as a tetramer (figure 99).

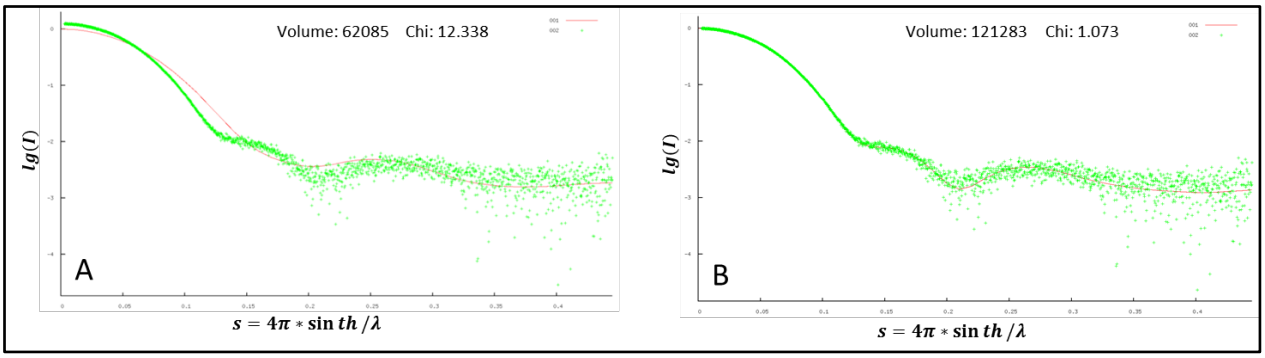


Figure 99: SAXS data fitting. A calculated scattering curve of a molecule (red line) is fitted with experimental measured data (green). If PA5507 is assumed to be a stable dimer in solution (14A) the curve does not match the data and the fitting value chi is far off from ideal. The tetramer assumption (14B) fits to the experimental data with a chi value of 1.073.

To exclude a concentration dependent oligomerization, five individual datasets were collected with different protein concentrations. Even at very low concentration the stable tetrameric complex is the only species that can be detected indicated by the typical scattering curve and the data quality resulting in chi values close to the ideal value of 1 (figure 99). The calculated SAXS envelope has a volume of 121280 Å³ in which a tetramer of PA5507 was fitted (figure 100).

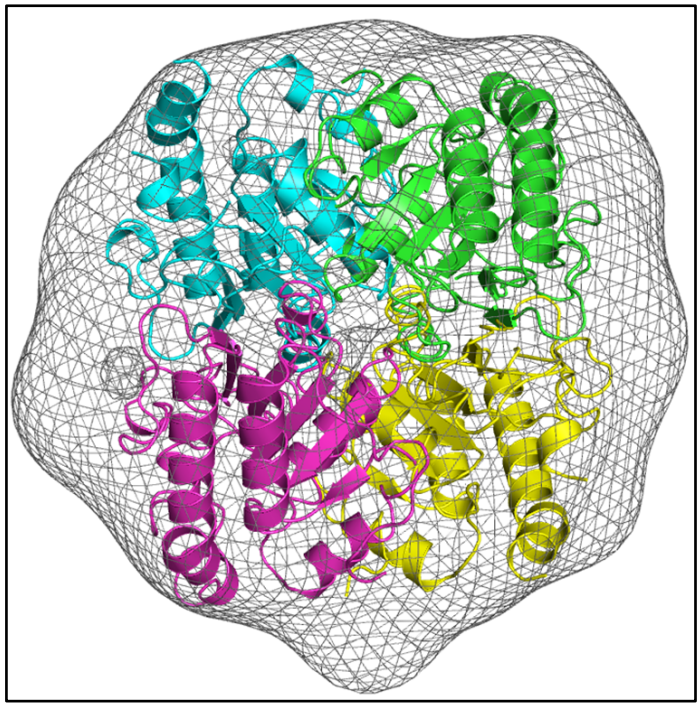


Figure 100: The calculated SAXS envelope fits the tetrameric PA5507.

The stable tetramer assembly (AB and CD) is realized by an interface, mostly established by polar contact of a small alpha helix, containing residues 82 to 93, one tyrosine residue (Y136) and a few hydrophobic interactions. The overall structure is shown in figure 101.

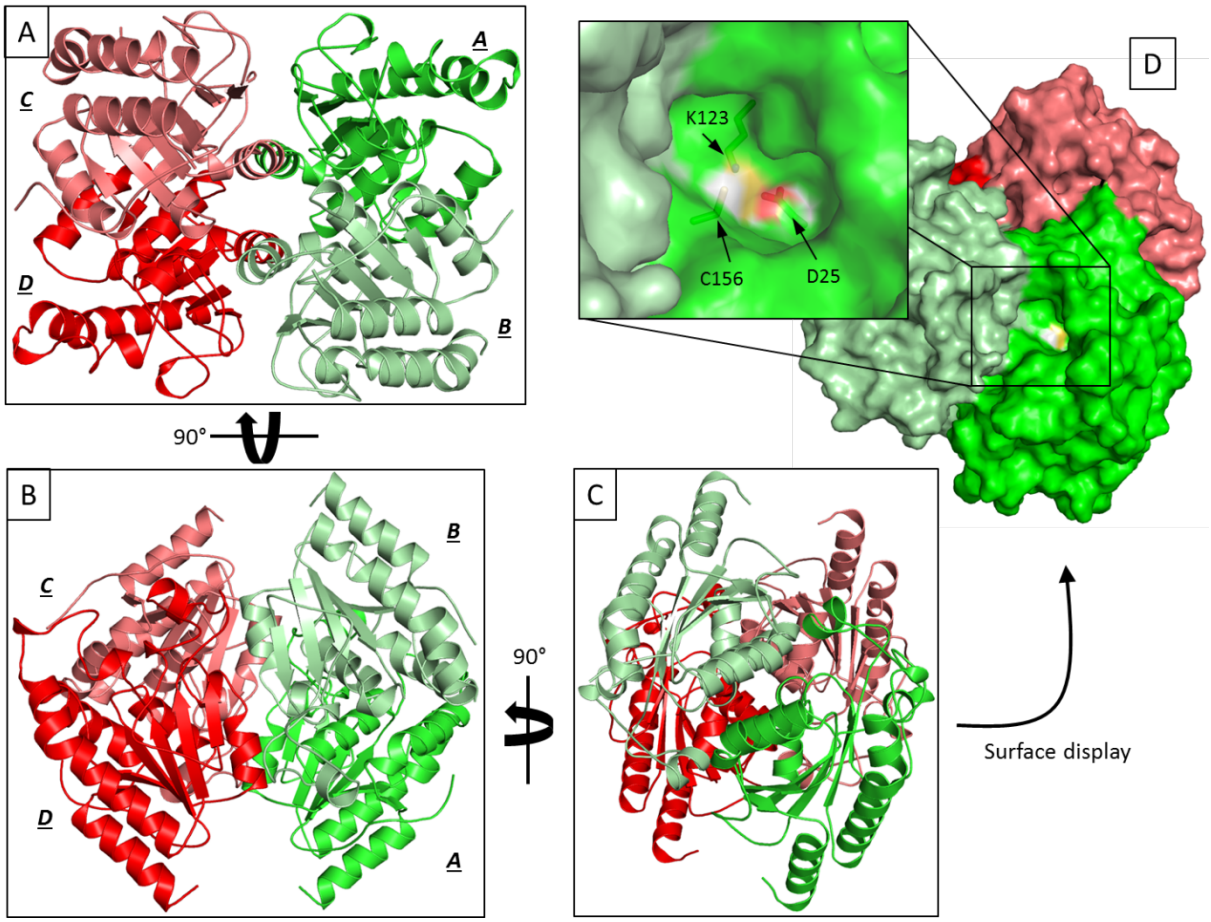


Figure 101: PA5507 is a homo-tetrameric protein with identical folded chains. Each dimer is represented in green and red with the monomers of the dimer is different sub-colors. The active pocket is depicted in D with a close up and the representation of the catalytic triad.

The four amino acid chains are almost of identical fold with an r.m.s.d. of 0.098 Å to each other, respectively (figure 8A). Due to the high flexibility, this regions linker was not completely traceable in both subunits and could not be modeled in chain B due to the lack of electron density. Super positioning of individual chains reveals that this loop region is closing a cavity that contains the catalytic triad involving side chains of a cysteine, aspartic acid and a lysine residue (figure 102B).

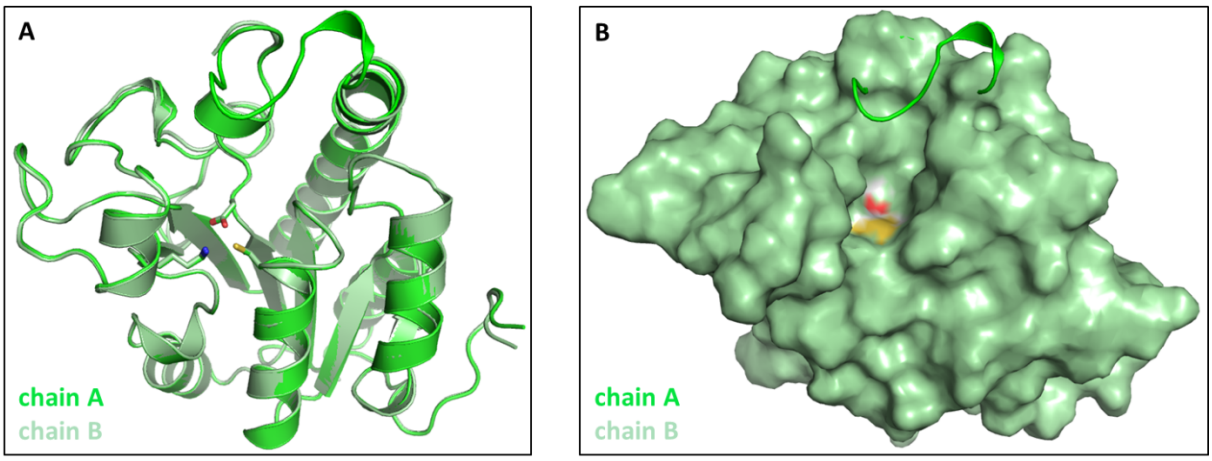


Figure 102: Superposition of chain A and chain B of PA5507 reveals an extra loop closing the active site. The active site is depicted in sticks (A).

Closer inspections of the loop region revealed an aspartic acid and a histidine residue separated by a proline, which allows the neighboring residues to be in close proximity with a distance of about 2.5 Å. This diade allows stabilization of positive charges within close proximity and is positioned at the top of the active site pocket (figure 103).

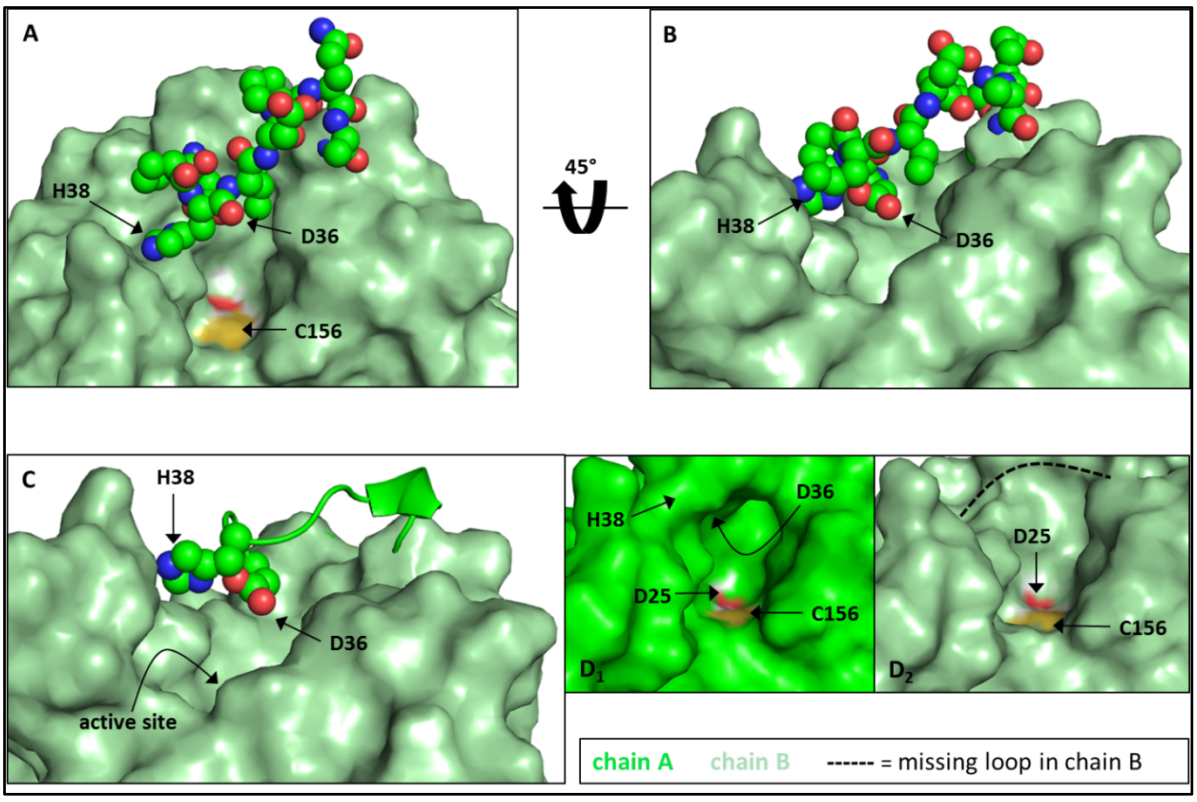


Figure 103: Active pocket is closed by the loop region containing amino acids 36 to 43 (A). This arrangement creates a diade of aspartic acid and histidine with a distance of 2.8 Å (B&C) to potentially stabilize positive charges.

This loop region seems to be highly flexible as indicated by high temperature factors (figure 104C). The closing of the active site might be the key to catalysis thus this can be only hypothesized since no data could be collected regarding the loop's flexibility.

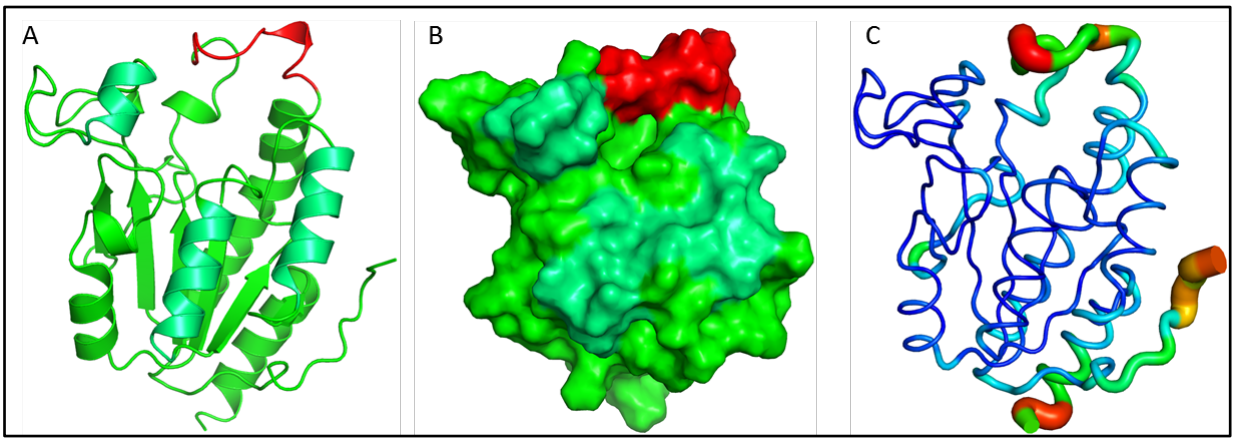


Figure 104: PA5507 monomer (A) the flexible part completing and closing the active site is shown in red (10A and 10B). A B-factor shows that PA5507 in general is a very rigid molecule except the loop region and the N- and C-termius to certain extend (10C).

5.5.4. Homology and active site analysis

In PA5507 the cysteine 156 is predicted to be the active nucleophile. It is located at the N-terminus of an alpha helix and can therefore be easily deprotonated into the anion state, which is supported by lysine 122, acting as a general base. This hypothesis is strongly supported by our crystallographic data. During refinement additional electron density which could not be interpreted by the polypeptide chain was observed, suggesting a covalently attached modification at the C156 which can neither be explained by oxidation of the sulfur nor a disulfide bond to a reductive agent. According to the size of the density, a ligand that covers a minimum number of seven atoms is proposed (figure 105).

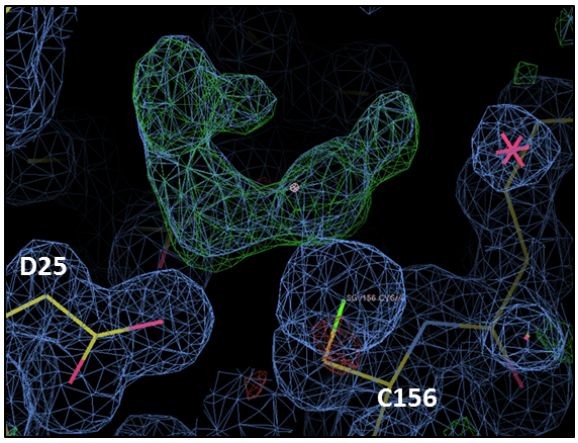
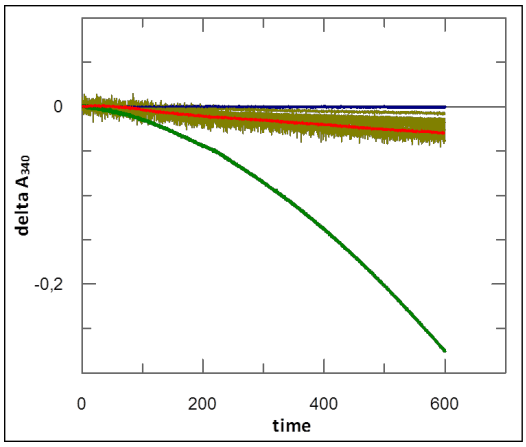


Figure 105: PA5507 chain A modification at the active site residue Cys156. Contour level of F_o-F_c is 1σ and $2F_o$-F_c = 3σ.

Results from structural searches selecting only cysteine hydrolases, PDBeFold and DALI both indicated a cysteine hydrolase from *Pseudomonas syringae*, (pdb: 3IRV) as most homologous structure apart form nicotinamidases that are also revealed as close homologous members. This protein family, however, involves a metal dependent ligand binding either using zinc (pdb: 2wt9), (Fyfe et al., 2009); 2h0r, (Hu et al., 2006); 3o92, (French et al., 2010); 3s2s, (Liu et al., 2011) or iron (pdb: 3pl1, (Petrella et al., 2010), respectively. Because PA5507 lacks the histidine residues involved in metal binding, a metal independent reaction mechanism must be employed and therefore PA5507 can be excluded from this family. Third a characterized N-carbamoyl sarcosine amidohydrolase (pdb: 1NBA, (Romão et al., 1992)) was identified by DALI search. Active site residues are located at but the surrounding active site environment is different. The N-carbamoyl sarcosine hydrolyzing enzyme provides polar contacts with close proximity of the active cysteine. The active site of PA5507, however, is clearly divided in two parts, the side close to the catalytic triade is lined with mostly polar residues while the other is of highly hydrophobic environment created by valine (V24, 128, 148, 151, 152, 157, 160), alanine (A159, 163, 164), leucine (L22, 173, 189) and isoleucine (I78) side chains.

5.5.5. Initial activity tests with amide substrates

PA5507 was tested for general amidase activity using various substrates. The employed modified glutamate dehydrogenase enzyme-coupled assay (Smith et al., 2009) did not reveals any substrate amongst the tested which was converted by PA5507 (figure 106).



Amide substrates (brown lines):

- Naphtylacetamide
- Salicylamide
- Propionamide
- Nicotinamide
- N-3Ac-4-hydroxyphenyl-butenamide
- Hexanoamide
- Butyramide
- Allantoin
- Acetamide
- N-Ac-anthranilate

- green: positive control
- red: negative control
- blue: buffer only

Figure 106: Enzyme-coupled assay to test amidase activity. The blue line represents a 'buffer-only' control; the red line shows the absorption of a negative control including everything but an amide. As positive control a known nicotinamidase and as substrate nicotinamide were used (green).

5.5.6. Structure in context

PA5507 is as member of the *qap*-operon co-transcribed with a characterized glutamine synthetase homolog (Ladner et al., 2012) and a putative N-formyl glutamate amidohydrolase, PA5509. The operon was suggested to be metabolizing either PQS (figure 14), an important *Pseudomonas* specific secondary metabolite or one of its precursor molecules (Tipton et al., 2013).

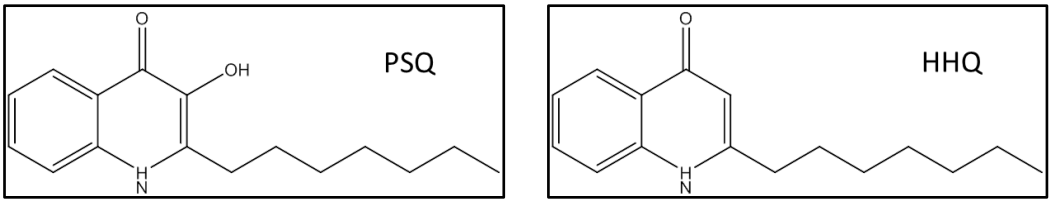


Figure 107: Chemical structure of PQS (*Pseudomonas* quinolone signal) and its direct chemical precursor molecule HHQ (2-heptyl-4-quinolone).

To elucidate this hypothesis, PA5507 was either co-crystallized with available precursor molecules of apo crystals of it were soaked with high concentration of the substances in order to trap only of the hypothesized ligands in the active site pocket (data not shown). Unfortunately, in none of the structures any of these ligands could be identified.

Concluding, the structure of PA5507, a member of the cysteine hydrolase superfamily could be solved in this study. An unidentified covalent modification of the active site cysteine was identified, however, the ligand itself remains unclear. Mass spectrometry analysis of the dissolved modified protein crystal did not reveal any ligand bound which

might be due to the harsh ionization method or the modification itself might be instable in solution. Therefore, the proteins substrate specificity remains unidentified.

6. Summary

Pseudomonas aeruginosa is a versatile opportunistic human pathogen with a rising number of antibiotic multidrug resistant strains. It can cause severe infections, especially to individuals in hospital environments where inappropriate therapy readily selects antibiotic resistant strains against which only a minor number of efficient agents are available. With this, the need of novel drug targets is rapidly emerging. Identification is tightly associated with knowing and understanding the bacteria's mode-of-action. As the first step, knowledge about the genomic composition was already achieved more than a decade ago, but understanding of the genome is still not accomplished. With about 40% of the genome still uncharacterized, it becomes clear that a plethora of new discoveries that might be used as new drug targets must be hidden within the large number of uncharacterized gene products.

In this work, results achieved during the pilot-phase of a new structure based approach with follow up metabolomic analysis, to elucidate previously uncharacterized operons, are presented. Two operons were selected, both comprising four previously uncharacterized genes. Selected gene clusters PA1621-PA1624 and PA5506-PA5509 which was termed "*qap*-operon" (Tipton et al., 2013) were structurally elucidated at the single gene level. Out of these genes, five crystals structures of proteins with unknown molecular function were determined at high resolution employing x-ray crystallographic methods. While this study was carried out, a sixth structure, that of PA5508 was reported by another research group (Ladner et al., 2012).

Structural characterization of PA1622 revealed a protein with an alpha/beta-fold (PA1622) comprising an active site featuring a classical serine peptidase triad, with an overall fold very similar to a previously reported thioester. During refinement of the structure, additional electron density was identified in the active site pocket, indicating a ligand molecule covalently bond to the active site serine. Due to weak signal for the ligand, an identification was not possible. However, a compound composed of a minimum number of 15 non-hydrogen atoms is proposed.

PA1623 was predicted to be a protein with glutathione-S-transferase activity. As the structure was determined it turned out that, instead of one glutathione molecule, the

protein had specifically bound two molecules of glutathione in the active site. Many structures have been reported that share the same typical GSH-fold as PA1623, yet this is the first one amongst the nu-class with a structurally resolved C-terminus, which might be involved in regulation of the protein's activity. The ligand bound structure indicated an involvement of the C-terminus as it restricts the access to the putative active sulfur atom of the bound GSH molecule.

The third determined structure is that of the periplasmic protein PA1624. Amongst all structures deposited to the *PDB*, no similar fold could be identified which renders this protein unique.

Genes coded in the '*qap*-operon' were shown to have a negative effect on concentration levels of PQS, an important second messenger involved in quorum sensing in *Pseudomonas* (Tipton et al., 2013). This operon is regulated by PA5506, a transcription regulator whose structure was solved in this study.

The tetrameric protein PA5506 was identified as a member of the RpiR-like transcription regulator family. Two distinct structures of the protein could be solved. In order to bind the metal, the protein undergoes a conformational within the C-terminus to complete the metal binding site. It was shown that a reorganization of the 'Tyr/Arg-switch loop' is necessary to complete the putative active site. Comparison to structural homologs, co-crystallized with their phosphorylated substrates, revealed a conserved phosphate position in the active site that occupied an identical position. This suggests that PA5506 also binds phosphorylated compounds, and in particular phosphorylated sugar compounds as it shows strong structural as well as moderate sequence homology to members of the sugar isomerase super family.

PA5507 was predicted as a protein with homology to members of the cysteine-hydrolase family. The structure revealed a protein with an amidase fold employing a His-Cys-Asp catalytic triad. Closer inspections of the active site identified uninterpretable electron density, which was assigned to a ligand that was covalently attached to the active site cysteine. Mass spectrometry experiments did not allow an identification of the bound ligand. The protein shows strong structural homology to both isochorismatases as well as amidases. A distinct substrate was not identified.

Pseudomonas aeruginosa PAO1 knockout strains, lacking a full operon were generated as well as gain of function mutants that harbor arabinose inducible plasmids coding for the full-length operon. Finally, an extraction method yielding high quality metabolite extract was established, leading to future experiments employing high-resolution mass spectrometry coupled with HPLC analysis.

7. Outlook

Five out of eight genes could be structurally characterized in this study, two of the *qap*-operon (Tipton et al., 2013) and three coded on the second selected operon. Hence, structural determination of the proteins PA5509 and PA1621 would be the next step in order to complete both operons.

PA5506, the transcription regulator of the *qap*-operon, already partially characterized would be the starting point for further experiments to first identify the regulating ligand, a putative phosphorylated sugar. First, metal binding of the protein should be verified by using for example calorimetric methods, also photometric methods like fluorescence titration could be used since at least four tyrosine residues are changing their chemical environment upon hypothetical metal binding as revealed in this study. Furthermore, once this question is answered experiments to identify its ligand are suggested to be carried out, employing first binding studies in the presence and absence of zinc to identify potential binds ligands. This is then followed by x-ray crystallographic methods by attempting either soaking or co-crystallization with the putative previously identified ligand. Third, if the ligand is identified, the problem of protein degradation should be elucidated. Tipton and coworker identified a DNA stretch which PA5506 seems to selectively recognize (Tipton et al., 2013). This oligonucleotide fragment is suggested to be used in co-crystallization attempts, because crystallization in complex with DNA might prevent the N-terminal domain from degradation.

However, the main goal of further studies would be the assignment of substrate and product for each operon. Therefore, metabolite extracts of generated operon knockout mutants need to be analyzed in a comparative metabolomic approach. Observed differences in metabolome composition amongst knockout *vs.* wild type *vs.* gain-of-function mutant might lead to a putative substrate/product pair, metabolized by the operon, and a general function assigned. Combining this with structural knowledge of all single gene products with, a functional hypothesis for every protein could be proposed. Appropriate biochemical experiments need to be carried out to prove proposed hypothetical functions for every protein as well as the operon.

8. Acknowledgement

This work was accomplished between April 2010 and February 2015 at the Max-Planck-Institute for Molecular Physiology in Dortmund, the University of Bayreuth as well as the Helmholtz-Centre for Infection research under the guidance of Prof. Dr. Wulf Blankenfeldt.

During the last five years of study at three different places, I constantly met great people, who contributed directly or indirectly to this work, without whom I could not have accomplished my doctoral degree. Herewith I would like to gratefully acknowledge:

Prof. Dr. Wulf Blankenfeldt as my mentor, for the opportunity to start this new project, his constant trust in my work, support of my 'crazy ideas' and constructive scientific guidance;

Prof. Dr. Michael Steinert for being my second supervisor and member of my examination committee;

Prof. Dr. Michael Hust for taking the chair in my examination committee;

Prof. Dr. Dimitri Mavrodi and **Prof. Linda Thomashow** for the chance to visit their labs at WSU and accomplish parts of my work under their guidance;

All my colleagues at *HZI* especially Allegra, Christina, Tobias aka 'Bockto', Nina, Moni, 'Flow', Vanessa, Juliane, Mumdooh and all that are not explicitly mentioned for constant support and an inspiring atmosphere;

All my former colleagues from the *MPI* in Dortmund, especially Mihai, Li, Shen, Petra, Tina and Melina; from *University Bayreuth* especially Philipp, 'Papa Weyand', Sébastien und 'der Franzose', Melanie, Frank, Norbeeeert, Susanne, Gabi und Renate; for material as well as administrative support, lots of 'spontaneous barbecue' and the amazing time;

My friends especially Stephan, Nadine, Philipp, Kai and all not explicitly mentioned – you altogether prevented me from going nuts…, hopefully.

and last but not least

My parents and my siblings for all you've done for me.

9. Appendix

9.1 Additional list of primers used in this study

Nr.	Oligoname	Sequenz (5' -> 3')	OD	µg	nmol	Konzentration [pmol/µl]	Vol. für 100pmol/µl	Tm [°C]	MW [g/mol]	GC-Gehalt
1	T4lyso NdeI_for	TAT TAC ATA TGA ATA TAT TTG AAA TGT TAC (30)	14.3	372	40,3	-	403	54,5 49,2	9217	16,7 %
2	T4lyso NdeI_rev	ATA TAA CAT ATG TAG ATT TTT ATA CGC GTC CC (32)	10.6	291	29,8	-	298	61,8 57,4	9772	31,3 %
3	T4Lyso stopXho_r	ATA TAA CTC GAG TTA TAG ATT TTT ATA CGC GTC CC (35)	8.5	236	22,1	-	221	64,8 61	10694	34,3 %
4	T4lyso_QC_for	CTT TAT TTT CAG GGC CAT ACG AAT ATA TTT GAA ATG TTA C (40)	18.3	502	40,9	-	409	65,3	12273	30 %
5	T4lyso_QC_rev	GTA ACA TTT CAA ATA TAT TCG TAT GGC CCT GAA AAT AAA G (40)	16.4	428	34,7	-	347	65,3	12310	30 %

Nr.	Oligoname	Sequenz (5' -> 3')	OD	µg	nmol	Konzentration [pmol/µl]	Vol. für 100pmol/µl	Tm [°C]	MW [g/mol]	GC-Gehalt
1	1621_for_Nde	AATATCATATGTCGCAAT CGGTTTTCTTC (29)	6.7	191	21,6	-	216	61,0	8832	34,5 %
2	1621_rev_Xho	AATTACTCGAGTCATGCC GAACGCTCCTC (29)	3.6	104	11,8	-	118	68,1	8797	51,7 %
3	1622_for_Nde	AATATCATATGAGCCTGC AGGTCGAGGAAG (30)	6.8	180	19,3	-	193	66,8	9304	46,7 %
4	1622_rev_Xho	TTATACTCGAGTCAGCGC GCGAAGAAGGC (29)	5.4	148	16,5	-	165	69,5	8951	55,2 %
5	1624_full_Nde_f	AATATCATATGCGAGGGT TCCTGTTGCTATC (31)	3.8	108	11,3	-	113	65,5	9516	41,9 %
6	1624_d18_nde_f	AATATCATATGGCCGACC TTCCCGGCAGCC (30)	5.3	153	16,8	-	168	70,9	9111	56,7 %
7	1624_rev_Xho	TTATACTCGAGTCAACGC AGCCAGTTGAG (29)	7.1	196	22,0	-	220	66,7	8901	48,3 %

Marker-specific primers and PCR programs

1) Gentamycin-specific primers
Primer: GM_UP (5' GGT GGC TCA AGT ATG GGC ATC A 3') 22-mer T_m=58.4 °C
Primer: GM_LOW (5' ATA GAG AGC CAC TGC GGG ATC G 3') 22-mer T_m=59.2 °C
Product length: 414 bp; 56.0% GC
PCR program: GM-UP-LW
94 °C 1' → 30 cycles of: 94 °C 35''
57 °C 20''
72 °C 25''

2) *sacB*-specific primers
Primer: SAC1 (5' GAT GTT TTC TTG CCT TTG ATG TTC 3') 24-mer T_m=67.4 °C
Primer: SAC2 (5' GTC TTT GCA TTA GCC GGA GAT C 3') 22-mer T_m=68.0 °C
Product length: 1050 bp
PCR program: SAC1-2
94 °C 1' → 30 cycles of: 94 °C 35''
53 °C 20''
72 °C 1'25''

3) Tet^r-specific primers
Primer: TET_UP (5' AGC GGT CCA GTG ATC GAA GTT A 3') 22-mer T_m=55.3 °C
Primer: TET_LOW (5' CAG GAG TCG CAT AAG GGA GAG C 3') 22-mer T_m=56.8 °C
Product length: 515 bp: 61.4% GC
PCR program: TET-UPLW
94 °C 1' → 30 cycles of: 94 °C 30''
52 °C 20''
72 °C 35''

4) Kan^r-specific primers
Primer: Kan_up (5' TGG CAA GAT CCT GGT ATC GGT 3') 21-mer T_m=68.9 °C
Primer: Kan_low (5' GAA ACA TGG CAA AGG TAG CGT 3') 21-mer T_m=66.8 °C
Product length: 530 bp: 45.1% GC
PCR program: GFPMUT3
94 °C 1' → 30 cycles of: 94 °C 30''
55 °C 20''
72 °C 35''

5) Cm^r-specific primers
Primer: Cm_up (5' ATC CCA ATG GCA TCG TAA AGA 3') 21-mer T_m=66.9 °C
Primer: Cm_low (5' AAG CAT TCT GCC GAC AT 3') 17-mer T_m=57.9 °C
Product length: 577 bp; 44.7% GC
PCR program: GFPMUT3

6) GFP-specific primers
Primer: gfpmut_up (5' CAT GGC CAA CAC TTG TCA CTA 3') 21-mer T_m=63.8 °C
Primer: gfpmut_low (5' CGA AAG GGC AGA TTG TG 3') 17-mer T_m=58.9 °C
Product length: 458 bp; 38.6% GC
PCR program: GFPMUT3

7) Amp^r-specific primers
Primer: bla1 (5' GGC CCC AGT GCT GCA ATG ATA C 3') 22-mer T_m=73.5 °C
Primer: bla2 (5' GAG TAT TCA ACA TTT CCG TGT CGC 3') 24-mer T_m=70.2 °C
Product length: 744 bp
PCR program: bla1-2
94 °C 1' → 30 cycles of: 94 °C 30''
57 °C 20''
72 °C 50''

9.2 Data collection statistics

Table 43: Data collection statistics; for selenomethionine data Friedel pairs were treated as separate observations. [#]Data were collected from single crystal; [+]Values in parentheses refer to the highest resolution shell. [α]Data collected on MXBL14.2/14.2, BESSY, Berlin (Mueller et al., 2012) [⌂]Data collected at home source; [β]Data collected at ESRF, ID23 (Nurizzo et al., 2006),[↕]Data collected on X10SA at SLS; [˚]Mosaicity and CC(1/2) reported by XDS (Kabsch, 2010; Karplus & Diederichs, 2012).

Dataset	Wavelength	Resolution range	Space group	No. total reflections	Completeness (%)	Redundancy / anno	<I/σ(I)>	R_{pim} (%)	CC(1/2)˚	Mosaicity˚	Notes
PA5506 α	0.91840	43.72-1.99 (2.05-1.99)	$P4_12_12$	217610 (18068)	100 (100)	7.2 (7.4)	20 (2)	2.7 (39.9)	0.99 (0.68)	0.10	native
PA5506_pk α	0.97981	48.35-2.2 (2.27-2.2)	$P2_12_12_1$	1033355 (91236)	100 (100)	26.6 (27.6) / 13.9 (14.1)	17.8 (4.0)	3.2 (23.9)	0.99 (0.92)	0.07	SeMet pk
PA5506_ip α	0.979927	48.36-3.2 (3.42-3.2)	$P2_12_12_1$	158170 (27498)	100 (100)	12.3 (12.1) / 6.6 (6.4)	12.0 (4.4)	6.2 (25.9)	0.99 (0.80)	0.17	SeMet ip
PA5506_pk α	0.979531	47.76-1.87 (1.91-1.87)	$P4_12_12$	432279 (28767)	100 (100)	12.7 (13.3) / 6.6 (6.8)	18.5 (2)	3 (41.4)	0.99 (72.3)	0.13	SeMet pk
PA5507 ⌂	1.5418	20-1.66 (1.70-1.66)	$P2_12_12$	172719 (6514)	96.9 (82.3)	3.7 (3.1)	17.35 (2.58)	3.5 (31.3)	0.99 (0.71)	0.467	native inhouse
PA5507 β	0.9724	42.13-1.36 (1.38-1.36)	$P2_12_12$	559039 (25882)	99.8 (99.8)	7.0 (6.5)	13.7 (2.3)	4.2 (44.6)	0.99 (0.72)	0.285	native
PA1622_pk α	0.9760	19.93-2.8 (2.92-2.8)	$P6_1$	895996 (109324)	99.7 (99.9)	14.8 (4.2) / 11.8 11.8	14.8 (4.2)	6.1 (20.5)	0.99 (0.89)	0.11	SeMet pk
PA1622 α	0.91840	45.57-2.10 (2.14-2.10)	$P6_1$	790175 (39646)	100 (100)	8.6 (8.8)	14.8 (2.2)	3.5 (38.6)	0.99 (0.70)	0.06	native
PA1623_pk ↕	0.97790	45.33-2.0 (2.05-2.0)	$P2_12_12$	193004 (15699)	99.9 (100)	6.4 (6.8) / 3.2 (3.4)	15.6 (2.9)	3.1 (23.1)	0.99 (0.94)	0.14	SeMet pk
PA1623_GSSG α	0. 91840	44.69-1.30 (1.33-1.30)	$P2_12_12_1$	1478145 (72620)	100 (100)	8.4 (8.3) / 4 (4)	14.3 (4)	5.1 (43.8)	0.998 (0.81)	0.12	Native GSSG
PA1623_apo α	0.91841	48.29-1.65 (1.75-1.65)	$P2_12_12$	212901 (34058)	99.7 (99.6)	6.25 (6.2)	18.55 (3.71)	3.0 (34.5)	0.99 (0.85)	0.18	Native
PA1624 α	0.979531	47.49-1.96 (2.01-1.96)	$P2_12_12_1$	3341268 (77395)	99.6 (98.4)	90.7 (29.5) / 46.9 (15.1)	23.6 (2.3)	3.4 (44.4)	0.99 (0.56)	0.16	SeMet pk
PA1624 α	0.91840	28.94-2.50 (2.6-2.5)	$P2_12_12_1$	69103 (8259)	97.8 (73)	3.7 (3.9)	5.1 (1.8)	12.1 (41.4)	0.978 (0.73)	0.33	native

9.3 Structural prediction

Table 434: Results for all protein sequences using the Phyre prediction server (Kelley & Sternberg, 2009); the five best hits are listed descending from top for all five proteins.

PDB ID	Description	Sequence coverage	Sequence identity
PA5506		88%	
4IVN	transcriptional regulator	95%	24%
3SHO	transcriptional regulator, rpir family	60%	24%
3FXA	sis domain protein	58%	15%
2XHZ	arabinose 5-phosphate isomerase	61%	14%
3ETN	putative phosphosugar isomerase involved in capsule	60%	16%
PA5507		94%	
2FQ1	non-ribosomal peptide synthetase entb	98%	24%
1NBA	Isochorismatase-like hydrolases	91%	26%
1NF9	Isochorismatase-like hydrolases	94%	24%
3IRV	cysteine hydrolase pspph_2384	96%	28%
1J2R	Isochorismatase-like hydrolases	90%	21%
PA1622		95%	
1CR6	epoxide hydrolase	98%	19%
3I28	epoxide hydrolase 2	99%	18%
1AZW	Proline iminopeptidase-like	97%	18%
4QLO	homoserine o-acetyltransferase	96%	12%
4D9J	16nm tetrahedral protein cage containing non-haem	96%	12%
PA1623		95%	
3C8E	yghu, glutathione s-transferase homologue	94%	38%
4ECJ	glutathione s-transferase prk13972	91%	100%
4MZW	nu-class glutathione transferase	92%	43%
3GXO	gst-like protein yfcg - oxidoreductase	90%	47%
4IKH	Glutathione transferase – Ps.fluor. pf-5	90%	49%
PA1624		24%	
2L26	rv0899 from mycobacterium tuberculosis	23%	18%
1R1M	OmpA-like domain	23%	18%
2AIZ	OmpA-like	21%	17%
3S0Y	periplasmic domain of motb	21%	16%
2K1S	nmr structure of the folded c-terminal fragment of yiad from2 escherichia coli	23%	23%

The best predicted structures are shown in figure 108.

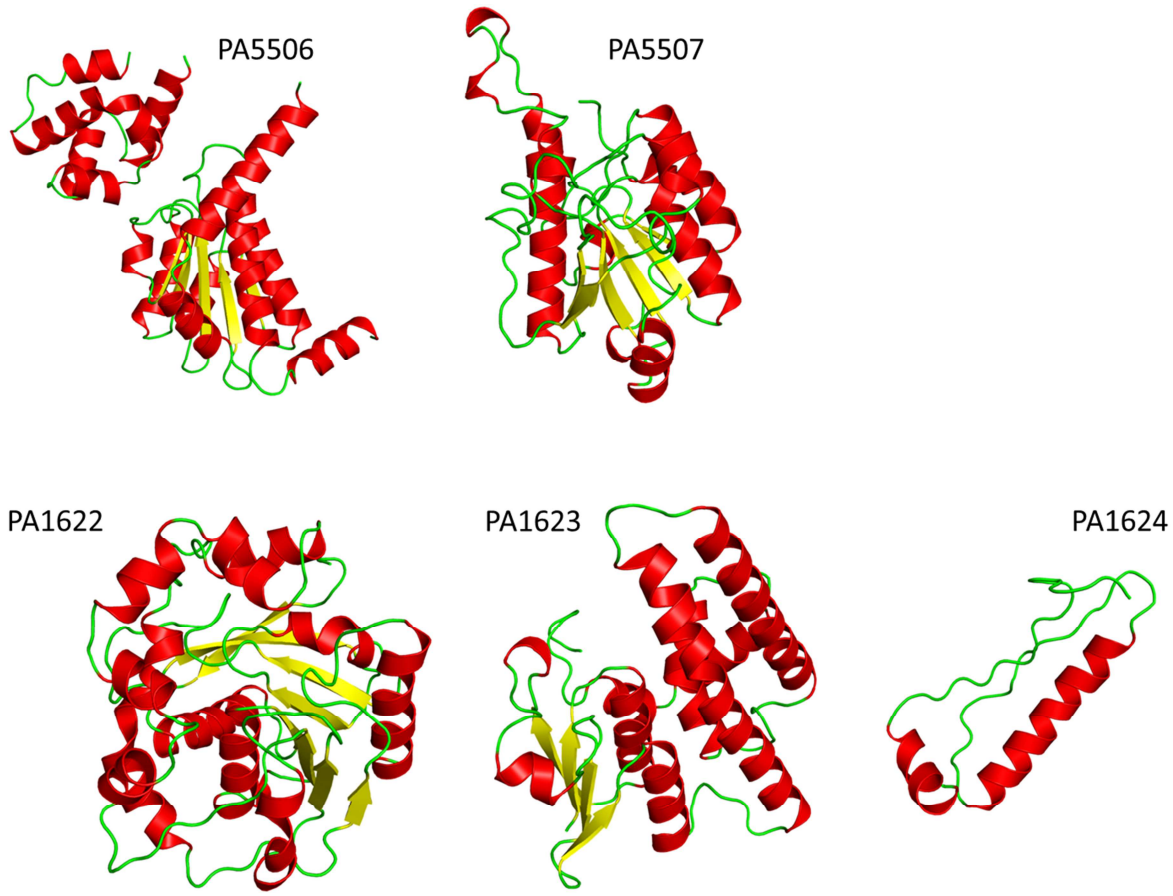


Figure 108: Structural prediction of investigated proteins in this study by phyre2.

9.4 Protein expression and purification

9.4.1 PA5506 expression and purification

The 285 amino acid protein PA5506 was expressed as an N-terminal His_6-tagged fusion protein. Typical affinity chromatography, tag cleavage and gel filtration yielded pure protein. A final yield of 4 mg of >95% pure native protein per liter of culture and 1 mg per liter in the case of selenomethionine labeled protein expression could be obtained. The protein elutes as a broad peak from an HP 16/60 S75 column.

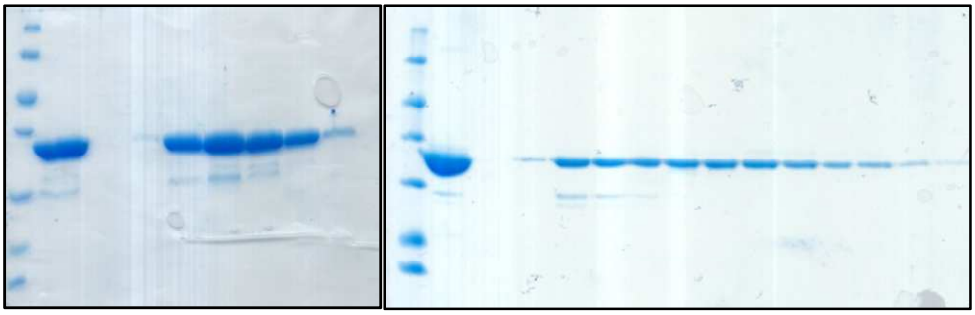


Figure 109: Left: SDS-PAGE analysis of SeMet PA5506 (32.3 kDa) after size exclusion chromatography, right native labeled protein. Marker includes proteins of 116, 66, 45, 35, 25, 18, 14 kDa size.

9.4.2 PA5507 expression and purification

PA5507, a 24 kDa protein was expressed as His_6-tagged SUMO fusion protein from a pOPINE vector (Berrow et al., 2007), cloned by DFP (http://www.dpf.mpi-dortmund.mpg.de/) (Oliner et al., 1993; Li & Elledge, 2007). Nickel-affinity chromatography, tag cleavage by SUMO-protease and size exclusion chromatography yielded 15 mg pure protein per liter of cell culture. PA5507 eluted from a Superdex 200 HP 26/60 as a single peak with a retention volume of 145 ml at peak maximum. The calculated mass of the protein is in the range of 100 to 110 kDa, indicating that the protein associates as tetramer in solution. About 15 mg of pure protein, according to Commassie stain were obtained from 1 liter cell culture. The protein was concentrated to 22 mg/ml prior to crystallization experiments.

9.4.3 PA5508 expression and purification

The glutamine synthetase (GS) homologue PA5508 was expressed from p10$ and purified by nickel affinity followed by typical tag cleavage. A yield of 6 mg/liter pure protein could be obtained after size exclusion chromatography. In order to elute PA5508 from the nickel column 500mM imidazole was supplied to the column. During gel filtration the protein eluted almost within the void volume as indicating a complex of large size.

9.4.4 PA1622 expression and purification

The putative enzyme PA1622 was expressed as N-terminal T7-lysozyme fusion from p10$. Nickel affinity chromatography and gel filtration yielded 12 mg of >95% pure protein per

liter cell culture in native conditions. 8 mg per liter were obtained when expressing the selenomethionine labeled protein.

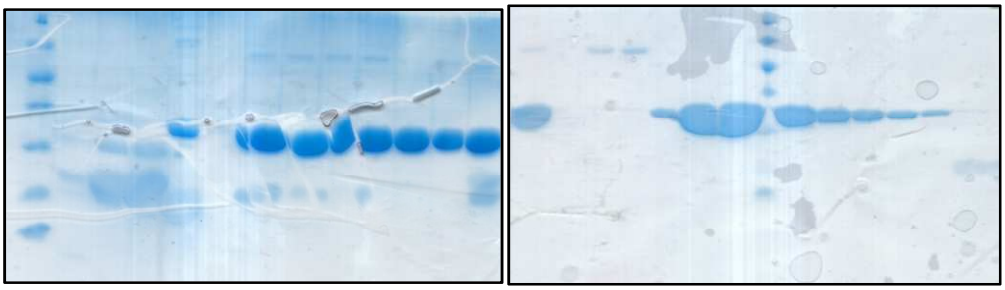


Figure 110: Left: SDS-PAGE analysis of native PA1622 (31.2 kDa) after size exclusion chromatography, right SeMet labeled protein. Marker includes proteins of 116, 66, 45, 35, 25, 18, 14 kDa size.

9.4.5 PA1623 expression and purification

The N-terminal GFP tagged PA1623 was purified by standard 2 step nickel affinity chromatography followed by size exclusion purification step. The finally obtained protein was >95% pure with a yield of 10 mg when expressing native protein and 8 mg of selenomethione labeled protein per liter of cultural broth.

9.4.6 PA1624 expression and purification

The predicted periplasmic protein was expressed from p10$ as an by signal peptide truncated (amino acid 1-18) N-terminal tagged fusion protein. Purified by metal affinity- and polished by size exclusion chromatography, 8 mg >95% pure native and 4 mg selenomethione labeled protein were yielded per liter cell culture.

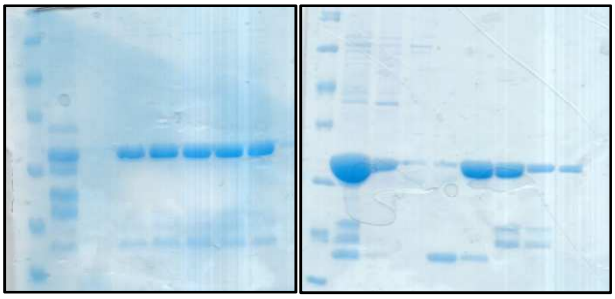


Figure 111: Left: SDS-PAGE analysis of native Δ_{18}PA1624 (27 kDa) after size exclusion chromatography, right SeMet labeled protein. Marker includes proteins of 116, 66, 45, 35, 25, 18, 14 kDa size.

9.4.7 Screening and optimization

From initial crystallization screen various conditions were identified to yield protein crystals. Several crystal shapes and sizes were obtained from which conditions yielding the best diffracting crystals were optimized in a two-dimensional grid screen manner. Improved size and shape was reached by varying the pH of buffer solution as well as the precipitant concentration of this condition. Thus, the protein concentration was also altered by setting up multiple drops within one condition. In table 45 are final mother liqour compositions summarized.

Table 44: Final mother liquor composition, protein concentration and cryo protecting agent

Protein	Mother liquor composition	Protein concentration	Cryo protection (mother liquor)
PA1622 native	0.24 M magnesium formate 22% PEG 3350	20 mg/ml	+ 20% PEG 400
PA1622 SeMet	0.1M HEPES pH7 5% PEG 6000	10 mg/ml	+ 20% glycerol
PA1623 native	0.1 M BisTris pH 6.5 0.1 M sodium potassium tartrate 22 % PEG 3350	15 mg/ml	+ 20% PEG400
PA1623 SeMet	0.1 M BisTris pH 6.7 0.1 M sodium potassium tartrate 18 % PEG 3350	14.5 mg/ml	+ 22% glycerol
PA1624 native	0.1 M MES pH 6 0.2 M Calcium acetate 20% PEG 8000	15 mg/ml	+ 20% glycerol
PA1624 SeMet	0.12 M tri-sodium citrate pH 5.5 0.2 M ammonium acetate 30% PEG 4000	20 mg/ml	+ 20% PEG 400
PA5506 native	0.2 M sodium chloride 10% PEG 3000 0.1 M sodium phosphate citrate pH 4.2	15 mg/ml	+ 22% PEG400
PA5506 SeMet	20% PEG 4000 0.2 M potassium fluoride 0.05 M KH2PO4	15 mg /ml	+ 20% PEG400
PA5507 native	0.26 M $CaCl_2$ 29% PEG 3350	10 mg / ml	+ 20% glycerol
PA5508 native	0.1 M MES pH 6.5 1.6 M Magnesium sulfate	4 mg/ml	+ 20% PEG 400

9.5 Fluorescence scan

To define the exact energy of the incident beam needed to match peak and inflection point of the protein incorporated selenomethinone atoms an x-ray fluorescence scan was performed before every experiment. An example is shown in figure 112.

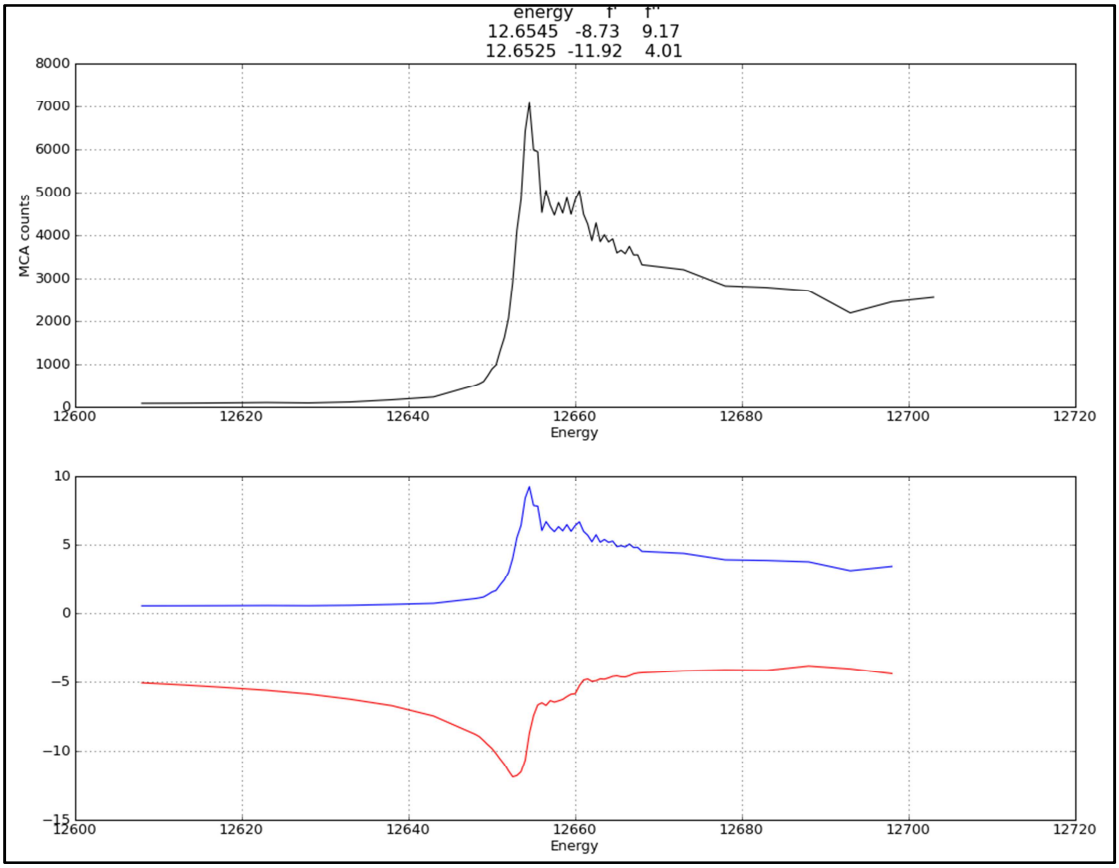


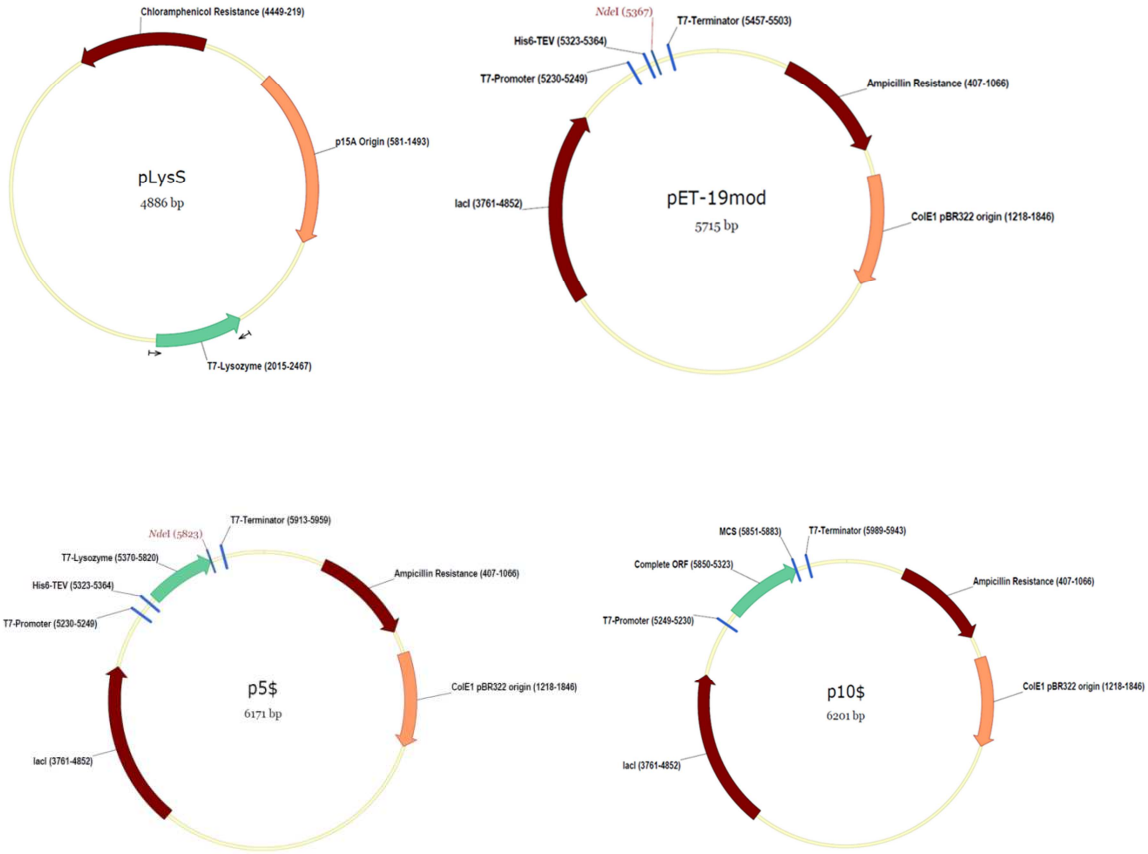
Figure 112: Result of a typical x-ray fluorescence scan to determine energy for peak and inflection point of selenium atoms

9.5.1 SAD phasing of PA5506/PA1622/PA1623/PA1624

Phases of labeled proteins were obtained by single or multi wavelength anomalous dispersion experiment carried out at the selenomethionine absorption edge. SHELXC (Sheldrick, 2010) were used to extract the anomalous signal and SHELXD (Schneider & Sheldrick, 2002) to locate heavy atom positions which were used as additional input for the program AutoSol (Adams et al., 2010; Echols et al., 2012) carrying out model building and initial refinement. This phase information was then transferred to native data sets if beneficial. Refined model was then used as search model for molecular replacement to

solve the structures in different space groups. The graphical interface hkl2map was used to visualize phasing results (Pape & Schneider, 2004).

9.6 Plasmid maps



10. References

Adair, F. W., Geftic, S. G., & Gelzer, J. (1969). *Appl. Microbiol.* **18**, 299–302.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000). *Science*. **287**, 2185–2195.

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., et al. (2010). *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221.

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., & Adams, P. D. (2012). *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367.

Allocati, N., Casalone, E., Masulli, M., Ceccarelli, I., Carletti, E., Parker, M. W., & Di Ilio, C. (1999). *FEBS Lett.* **445**, 347–350.

Allocati, N., Federici, L., Masulli, M., & Di Ilio, C. (2009). *FEBS J.* **276**, 58–75.

Aloush, V., Navon-Venezia, S., Seigman-Igra, Y., Cabili, S., & Carmeli, Y. (2006). *Antimicrob. Agents Chemother.* **50**, 43–48.

Altermann, W. & Kazmierczak, J. (2003). *Res. Microbiol.* **154**, 611–617.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.

Altschul, S. F., Wootton, J. C., Gertz, E. M., Agarwala, R., Morgulis, A., Schäffer, A. A., & Yu, Y.-K. (2005). *FEBS J.* **272**, 5101–5109.

Amouyal, M. & Buc, H. (1987). *J. Mol. Biol.* **195**, 795–808.

Apic, G., Gough, J., & Teichmann, S. A. (2001). *J. Mol. Biol.* **310**, 311–325.

Arber, W. (2014). *Life Basel Switz.* **4**, 217–224.

Argiriadi, M. A., Morisseau, C., Hammock, B. D., & Christianson, D. W. (1999). *Proc. Natl. Acad. Sci. U. S. A.* **96**, 10637–10642.

Aslund, F., Berndt, K. D., & Holmgren, A. (1997). *J. Biol. Chem.* **272**, 30780–30786.

Bagos, P. G., Nikolaou, E. P., Liakopoulos, T. D., & Tsirigos, K. D. (2010). *Bioinforma. Oxf. Engl.* **26**, 2811–2817.

Bassetti, M., Merelli, M., Temperoni, C., & Astilean, A. (2013). *Ann. Clin. Microbiol. Antimicrob.* **12**, 22.

Baxter, A., Chambers, M., Edfeldt, F., Edman, K., Freeman, A., Johansson, C., King, S., Morley, A., Petersen, J., Rawlins, P., et al. (2011). *Bioorg. Med. Chem. Lett.* **21**, 777–780.

Bendig, J. W., Kyle, P. W., Giangrande, P. L., Samson, D. M., & Azadian, B. S. (1987). *J. R. Soc. Med.* **80**, 316–317.

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). *J. Mol. Biol.* **340**, 783–795.

Bereket, W., Hemalatha, K., Getenet, B., Wondwossen, T., Solomon, A., Zeynudin, A., & Kannan, S. (2012). *Eur. Rev. Med. Pharmacol. Sci.* **16**, 1039–1044.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bharathi, S., Raman, G. V., Mohan, D. M., & Krishnan, A. (2014). *Indian J. Ophthalmol.* **62**, 958–960.

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). *PloS One*. **2**, e197.

Biswal, B. K., Morisseau, C., Garen, G., Cherney, M. M., Garen, C., Niu, C., Hammock, B. D., & James, M. N. G. (2008). *J. Mol. Biol.* **381**, 897–912.

Blackburn, A. C., Woollatt, E., Sutherland, G. R., & Board, P. G. (1998). *Cytogenet. Cell Genet.* **83**, 109–114.

Blankenfeldt, W. & Parsons, J. F. (2014). *Curr. Opin. Struct. Biol.* **29C**, 26–33.

Blommel, P. G., Martin, P. A., Seder, K. D., Wrobel, R. L., & Fox, B. G. (2009). *Methods Mol. Biol. Clifton NJ*. **498**, 55–73.

Bolivar, F., Rodriguez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L., Boyer, H. W., Crosa, J. H., & Falkow, S. (1977). *Gene*. **2**, 95–113.

Bond, C. S. (2003). *Bioinforma. Oxf. Engl.* **19**, 311–312.

Braman, J., Papworth, C., & Greener, A. (1996). *Methods Mol. Biol. Clifton NJ*. **57**, 31–44.

Braun, N. (2013). *Univ. Bayreuth*.

Buschbeck, M. & Ullrich, A. (2005). *J. Biol. Chem.* **280**, 2659–2667.

Butt, T. R., Edavettal, S. C., Hall, J. P., & Mattern, M. R. (2005). *Protein Expr. Purif.* **43**, 1–9.

Callahan, P. L., Mizutani, S., & Colonno, R. J. (1985). *Proc. Natl. Acad. Sci. U. S. A.* **82**, 732–736.

Calvin, N. M. & Hanawalt, P. C. (1988). *J. Bacteriol.* **170**, 2796–2801.

Chandonia, J.-M. & Brenner, S. E. (2006). *Science*. **311**, 347–351.

Chapman, R. D., Palancade, B., Lang, A., Bensaude, O., & Eick, D. (2004). *Nucleic Acids Res.* **32**, 35–44.

Chayen, N. E. & Saridakis, E. (2008). *Nat. Methods*. **5**, 147–153.

Cheetham, G. M., Jeruzalmi, D., & Steitz, T. A. (1999). *Nature*. **399**, 80–83.

Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2010). *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21.

Chen, W. P. & Kuo, T. T. (1993). *Nucleic Acids Res.* **21**, 2260.

Cherukuvada, S. L., Seshasayee, A. S. N., Raghunathan, K., Anishetty, S., & Pennathur, G. (2005). *PLoS Comput Biol*. **1**, e28.

Chiu, H.-J., Grant, J. C., Farr, C. L., Jaroszewski, L., Knuth, M. W., Miller, M. D., Elsliger, M.-A., Deacon, A. M., Godzik, A., Lesley, S. A., et al. (2014). *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2640–2651.

Choi, K.-H., Kumar, A., & Schweizer, H. P. (2006). *J. Microbiol. Methods*. **64**, 391–397.

Choi, K.-H. & Schweizer, H. P. (2005). *BMC Microbiol.* **5**, 30.

Chruszcz, M., Potrzebowski, W., Zimmerman, M. D., Grabowski, M., Zheng, H., Lasota, P., & Minor, W. (2008). *Protein Sci. Publ. Protein Soc.* **17**, 623–632.

Claxton, H. B., Akey, D. L., Silver, M. K., Admiraal, S. J., & Smith, J. L. (2009). *J. Biol. Chem.* **284**, 5021–5029.

Cohen, S. N., Chang, A. C. Y., Boyer, H. W., & Helling, R. B. (1973). *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3240–3244.

Cohen, S. N., Chang, A. C. Y., & Hsu, L. (1972). *Proc. Natl. Acad. Sci. U. S. A.* **69**, 2110–2114.

Cole, T., Kathman, A., Koszelak, S., & McPherson, A. (1995). *Anal. Biochem.* **231**, 92–98.

Colovos, C., Cascio, D., & Yeates, T. O. (1998). *Struct. Lond. Engl. 1993*. **6**, 1329–1337.

Consortium, T. C. elegans S. (1998). *Science*. **282**, 2012–2018.

Cordingley, M. G., Callahan, P. L., Sardana, V. V., Garsky, V. M., & Colonno, R. J. (1990). *J. Biol. Chem.* **265**, 9062–9065.

Cornett, R., James, M. O., Henderson, G. N., Cheung, J., Shroads, A. L., & Stacpoole, P. W. (1999). *Biochem. Biophys. Res. Commun.* **262**, 752–756.

Dacheux, D., Goure, J., Chabert, J., Usson, Y., & Attree, I. (2001). *Mol. Microbiol.* **40**, 76–85.

Dacheux, D., Toussaint, B., Richard, M., Brochier, G., Croize, J., & Attree, I. (2000). *Infect. Immun.* **68**, 2916–2924.

Darch, S. E., McNally, A., Harrison, F., Corander, J., Barr, H. L., Paszkiewicz, K., Holden, S., Fogarty, A., Crusz, S. A., & Diggle, S. P. (2015). *Sci. Rep.* **5**,.

D'Argenio, D. A., Gallagher, L. A., Berg, C. A., & Manoil, C. (2001). *J. Bacteriol.* **183**, 1466–1471.

Davis, C., Carberry, S., Schrettl, M., Singh, I., Stephens, J. C., Barry, S. M., Kavanagh, K., Challis, G. L., Brougham, D., & Doyle, S. (2011). *Chem. Biol.* **18**, 542–552.

Davis, I. W., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2004). *Nucleic Acids Res.* **32**, W615–W619.

Van Delden, C. (2007). *Int. J. Antimicrob. Agents*. **30 Suppl 1**, S71–S75.

Dorward, D. W. & Garon, C. F. (1990). *Appl. Environ. Microbiol.* **56**, 1960–1962.

Downie, A. W. (1972). *J. Gen. Microbiol.* **73**, 1–11.

Droege, M. & Hill, B. (2008). *J. Biotechnol.* **136**, 3–10.

Emsley, P. & Cowtan, K. (2004). *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132.

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501.

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., et al. (2013). *G3 GenesGenomesGenetics*. **4**, 389–398.

Englard, S. & Seifter, S. (1990). *Methods in Enzymology*, Murray P. Deutscher, edited by, pp. 285–300. Academic Press.

Ettwig, K. F., Butler, M. K., Le Paslier, D., Pelletier, E., Mangenot, S., Kuypers, M. M. M., Schreiber, F., Dutilh, B. E., Zedelius, J., de Beer, D., et al. (2010). *Nature*. **464**, 543–548.

Evans, J. P., Meslin, E. M., Marteau, T. M., & Caulfield, T. (2011). *Science*. **331**, 861–862.

Evans, P. R. & Murshudov, G. N. (2013). *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204–1214.

Fernández-Cañón, J. M. & Peñalva, M. A. (1998). *J. Biol. Chem.* **273**, 329–337.

Finkelstein, A. V., Gutun, A. M., & Badretdinov AYa, null (1993). *FEBS Lett.* **325**, 23–28.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). *Nucleic Acids Res.* **39**, W29–W37.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., & Merrick, J. M. (1995). *Science*. **269**, 496–512.

French, J. B., Cen, Y., Sauve, A. A., & Ealick, S. E. (2010). *Biochemistry (Mosc.)*. **49**, 8803–8812.

Frengen, E., Weichenhan, D., Zhao, B., Osoegawa, K., van Geel, M., & de Jong, P. J. (1999). *Genomics*. **58**, 250–253.

Fröbel, J., Rose, P., Lausberg, F., Blümmel, A.-S., Freudl, R., & Müller, M. (2012). *Nat. Commun.* **3**, 1311.

Froger, A. & Hall, J. E. (2007). *J. Vis. Exp. JoVE*.

Fyfe, P. K., Rao, V. A., Zemla, A., Cameron, S., & Hunter, W. N. (2009). *Angew.Chem.Int.Ed.Engl.* **48**, 9176 – null.

Garrity, J., Gardner, J. G., Hawse, W., Wolberger, C., & Escalante-Semerena, J. C. (2007). *J. Biol. Chem.* **282**, 30239–30245.

Gay, P., Le Coq, D., Steinmetz, M., Berkelman, T., & Kado, C. I. (1985). *J. Bacteriol.* **164**, 918–921.

Gay, P., Le Coq, D., Steinmetz, M., Ferrari, E., & Hoch, J. A. (1983). *J. Bacteriol.* **153**, 1424–1431.

GE (2010). *GE Healthc. Handb.* **1**, 7–12.

Gehret, J. J., Gu, L., Gerwick, W. H., Wipf, P., Sherman, D. H., & Smith, J. L. (2011). *J.Biol.Chem.* **286**, 14445–14454.

Gerlt, J. A., Allen, K. N., Almo, S. C., Armstrong, R. N., Babbitt, P. C., Cronan, J. E., Dunaway-Mariano, D., Imker, H. J., Jacobson, M. P., Minor, W., et al. (2011). *Biochemistry (Mosc.)*. **50**, 9950–9962.

Gonzales, M. F., Brooks, T., Pukatzki, S. U., & Provenzano, D. (2013). *J. Vis. Exp. JoVE*.

Goodwin, P. C. (2014). *Mol. Reprod. Dev.* n/a – n/a.

Gorbunov, A. A. & Skvortsov, A. M. (1995). *Adv. Colloid Interface Sci.* **62**, 31–108.

Gorbunov, A. & Trathnigg, B. (2002). *J. Chromatogr. A*. **955**, 9–17.

Green, S. K., Schroth, M. N., Cho, J. J., Kominos, S. K., & Vitanza-jack, V. B. (1974). *Appl. Microbiol.* **28**, 987–991.

Griffith, F. (1928). *J. Hyg. (Lond.)*. **27**, 113–159.

Di Guan, C., Li, P., Riggs, P. D., & Inouye, H. (1988). *Gene*. **67**, 21–30.

Haas, C. & Drenth, J. (1999). *J. Cryst. Growth*. **196**, 388–394.

Hakemi Vala, M., Hallajzadeh, M., Hashemi, A., Goudarzi, H., Tarhani, M., Sattarzadeh Tabrizi, M., & Bazmi, F. (2014). *Ann. Burns Fire Disasters*. **27**, 8–13.

Hammarström, M., Hellgren, N., van Den Berg, S., Berglund, H., & Härd, T. (2002). *Protein Sci. Publ. Protein Soc.* **11**, 313–321.

Hansen, L. H., Knudsen, S., & Sørensen, S. J. (1998). *Curr. Microbiol.* **36**, 341–347.

Harmer, N. J. (2010). *J. Mol. Biol.* **400**, 379–392.

Hauser, A. R., Kang, P. J., & Engel, J. N. (1998). *Mol. Microbiol.* **27**, 807–818.

Heiner, C. R., Hunkapiller, K. L., Chen, S. M., Glass, J. I., & Chen, E. Y. (1998). *Genome Res.* **8**, 557–561.

Hendrickson, W. A. (1991). *Science*. **254**, 51–58.

Heringa, J. & Taylor, W. R. (1997). *Curr. Opin. Struct. Biol.* **7**, 416–421.

Hewitt, L. & McDonnell, J. M. (2004). *Methods Mol. Biol. Clifton NJ*. **278**, 1–16.

Hickey, C. M., Wilson, N. R., & Hochstrasser, M. (2012). *Nat. Rev. Mol. Cell Biol.* **13**, 755–766.

Hidron, A. I., Edwards, J. R., Patel, J., Horan, T. C., Sievert, D. M., Pollock, D. A., Fridkin, S. K., National Healthcare Safety Network Team, & Participating National Healthcare Safety Network Facilities (2008). *Infect. Control Hosp. Epidemiol. Off. J. Soc. Hosp. Epidemiol. Am.* **29**, 996–1011.

Hoang, T. T., Karkhoff-Schweizer, R. R., Kutchma, A. J., & Schweizer, H. P. (1998). *Gene*. **212**, 77–86.

Hochstrasser, M. (2009). *Nature*. **458**, 422.

Hochuli, E., Döbeli, H., & Schacher, A. (1987). *J. Chromatogr.* **411**, 177–184.

Hofmann, M. A. & Brian, D. A. (1991). *BioTechniques*. **11**, 30–31.

Hogardt, M. & Heesemann, J. (2013). *Curr. Top. Microbiol. Immunol.* **358**, 91–118.

Holm, L. & Rosenström, P. (2010). *Nucleic Acids Res.* **38**, W545–W549.

Howard, E. I., Sanishvili, R., Cachau, R. E., Mitschler, A., Chevrier, B., Barth, P., Lamour, V., Van Zandt, M., Sibley, E., Bon, C., et al. (2004). *Proteins*. **55**, 792–804.

Hu, G., Taylor, A. B., McAlister-Henn, L., & Hart, P. J. (2006). *Arch.Biochem.Biophys.* **461**, 66–75.

Hwang, J., Kim, B. S., Jang, S. Y., Lim, J. G., You, D.-J., Jung, H. S., Oh, T.-K., Lee, J.-O., Choi, S. H., & Kim, M. H. (2013). *Proc. Natl. Acad. Sci. U. S. A.* **110**, E2829–E2837.

IDT, C., USA (2012).

Ingham, K. C. (1984). *Methods in Enzymology*, William B. Jakoby, edited by, pp. 351–356. Academic Press.

Ingham, K. C. (1990). *Methods in Enzymology*, Murray P. Deutscher, edited by, pp. 301–306. Academic Press.

Inoue, H., Nojima, H., & Okayama, H. (1990). *Gene*. **96**, 23–28.

Ishida, T. & Kinoshita, K. (2007). *Nucleic Acids Res.* **35**, W460–W464.

Ishida, T. & Kinoshita, K. (2008). *Bioinforma. Oxf. Engl.* **24**, 1344–1348.

Islam, S. A., Luo, J., & Sternberg, M. J. (1995). *Protein Eng.* **8**, 513–525.

Itah, A. Y., Brooks, A. A., Ogar, B. O., & Okure, A. B. (2009). *Bull. Environ. Contam. Toxicol.* **83**, 318–327.

Jensen, L. M., Walker, E. J., Jans, D. A., & Ghildyal, R. (2015). *Methods Mol. Biol. Clifton NJ*. **1221**, 129–141.

Jeruzalmi, D. & Steitz, T. A. (1998). *EMBO J.* **17**, 4101–4113.

Kabsch, W. (2010). *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132.

Kakutani, S. (1950). *Proc. Natl. Acad. Sci. U. S. A.* **36**, 319–323.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). *Nucleic Acids Res.* **42**, D199–D205.

Kapust, R. B., Tözsér, J., Copeland, T. D., & Waugh, D. S. (2002). *Biochem. Biophys. Res. Commun.* **294**, 949–955.

Kapust, R. B., Tözsér, J., Fox, J. D., Anderson, D. E., Cherry, S., Copeland, T. D., & Waugh, D. S. (2001). *Protein Eng.* **14**, 993–1000.

Karplus, P. A. & Diederichs, K. (2012). *Science*. **336**, 1030–1033.

Katti, S. K., LeMaster, D. M., & Eklund, H. (1990). *J. Mol. Biol.* **212**, 167–184.

Kelley, L. A. & Sternberg, M. J. E. (2009). *Nat. Protoc.* **4**, 363–371.

Klevens, R. M., Edwards, J. R., Richards, C. L., Jr, Horan, T. C., Gaynes, R. P., Pollock, D. A., & Cardo, D. M. (2007). *Public Health Rep. Wash. DC 1974*. **122**, 160–166.

Klockgether, J., Cramer, N., Wiehlmann, L., Davenport, C. F., & Tümmler, B. (2011). *Front. Microbiol.* **2**, 150.

Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J., & Svergun, D. I. (2003). *J. Appl. Crystallogr.* **36**, 1277–1282.

Kong, G. K.-W., Polekhina, G., McKinstry, W. J., Parker, M. W., Dragani, B., Aceto, A., Paludi, D., Principe, D. R., Mannervik, B., & Stenberg, G. (2003). *J. Biol. Chem.* **278**, 1291–1302.

Konhauser, K. O., Lalonde, S. V., Planavsky, N. J., Pecoits, E., Lyons, T. W., Mojzsis, S. J., Rouxel, O. J., Barley, M. E., Rosìere, C., Fralick, P. W., et al. (2011). *Nature*. **478**, 369–373.

Kozin, M. B. & Svergun, D. I. (2001). *J. Appl. Crystallogr.* **34**, 33–41.

Krall, R., Schmidt, G., Aktories, K., & Barbieri, J. T. (2000). *Infect. Immun.* **68**, 6066–6068.

Krissinel, E. (2007). *Bioinforma. Oxf. Engl.* **23**, 717–723.

Krissinel, E. (2010). *J. Comput. Chem.* **31**, 133–143.

Krissinel, E. (2011). *Acta Crystallogr. D Biol. Crystallogr.* **67**, 376–385.

Krissinel, E. & Henrick, K. (2004). *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2256–2268.

Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). *J. Mol. Biol.* **305**, 567–580.

Kusters, J. G., van Vliet, A. H. M., & Kuipers, E. J. (2006). *Clin. Microbiol. Rev.* **19**, 449–490.

Laborde, E. (2010). *Cell Death Differ.* **17**, 1373–1380.

Lack, N., Lowe, E. D., Liu, J., Eltis, L. D., Noble, M. E. M., Sim, E., & Westwood, I. M. (2008). *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **64**, 2–7.

Ladner, J. E., Atanasova, V., Dolezelova, Z., & Parsons, J. F. (2012). *Biochemistry (Mosc.).* **51**, 10121–10123.

Laemmli, U. K. (1970). *Nature*. **227**, 680–685.

Lakkis, C. & Fleiszig, S. M. (2001). *J. Clin. Microbiol.* **39**, 1477–1486.

Laskowski, R. A. & Swindells, M. B. (2011). *J. Chem. Inf. Model.* **51**, 2778–2786.

Lee, D. G., Urbach, J. M., Wu, G., Liberati, N. T., Feinbaum, R. L., Miyata, S., Diggins, L. T., He, J., Saucier, M., Déziel, E., et al. (2006). *Genome Biol.* **7**, R90.

Lemaitre, N., Callebaut, I., Frenois, F., Jarlier, V., & Sougakoff, W. (2001). *Biochem. J.* **353**, 453–458.

Lessnau, K.-D., Cunha, B. A., Dua, P., & Qarah, S. (2014). *Medscape Drugs Dis. Proced.*

Letunic, I., Doerks, T., & Bork, P. (2014). *Nucleic Acids Res.* gku949.

Levitt, M. & Chothia, C. (1976). *Nature*. **261**, 552–558.

Liebert, M. A. (2014). *Genet. Eng. Biotechnol. News*. **34**, 18–18.

Li, J., Shi, C., Gao, Y., Wu, K., Shi, P., Lai, C., Chen, L., Wu, F., & Tian, C. (2012). *J. Mol. Biol.* **415**, 382–392.

Li, S. J. & Hochstrasser, M. (1999). *Nature*. **398**, 246–251.

Lister, P. D., Wolter, D. J., & Hanson, N. D. (2009). *Clin. Microbiol. Rev.* **22**, 582–610.

Liu, H. & Naismith, J. H. (2008). *BMC Biotechnol.* **8**, 91.

Liu, X., Zhang, H., Wang, X.-J., Li, L.-F., & Su, X.-D. (2011). *PloS One*. **6**, e24227.

Livermore, D. M. (2002). *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **34**, 634–640.

Long, F., Vagin, A. A., Young, P., & Murshudov, G. N. (2008). *Acta Crystallogr. D Biol. Crystallogr.* **64**, 125–132.

Looman, A. C., Bodlaender, J., Comstock, L. J., Eaton, D., Jhurani, P., de Boer, H. A., & van Knippenberg, P. H. (1987). *EMBO J.* **6**, 2489–2492.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S., & Richardson, D. C. (2003). *Proteins*. **50**, 437–450.

Ludewig, H. (2012). *Univ. Bayreuth*.

Lyczak, J. B., Cannon, C. L., & Pier, G. B. (2000). *Microbes Infect. Inst. Pasteur*. **2**, 1051–1060.

Mahajan-Miklos, S., Tan, M. W., Rahme, L. G., & Ausubel, F. M. (1999). *Cell*. **96**, 47–56.

Mannervik, B., Alin, P., Guthenberg, C., Jensson, H., Tahir, M. K., Warholm, M., & Jörnvall, H. (1985). *Proc. Natl. Acad. Sci.* **82**, 7202–7206.

Mao, F., Dam, P., Chou, J., Olman, V., & Xu, Y. (2009). *Nucleic Acids Res.* **37**, D459–D463.

Mao, X., Ma, Q., Zhou, C., Chen, X., Zhang, H., Yang, J., Mao, F., Lai, W., & Xu, Y. (2014). *Nucleic Acids Res.* **42**, D654–D659.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., et al. (2011). *Nucleic Acids Res.* **39**, D225–D229.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., et al. (2013). *Nucleic Acids Res.* **41**, D348–D352.

Mardis, E. R. (2011). *Nature*. **470**, 198–203.

Martin, J. L. (1995). *Struct. Lond. Engl. 1993*. **3**, 245–250.

De la Mata, M. & Kornblihtt, A. R. (2006). *Nat. Struct. Mol. Biol.* **13**, 973–980.

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M., Rokas, A., Yandava, C. N., Engels, R., Zeng, E., et al. (2008). *Proc. Natl. Acad. Sci. U. S. A.* **105**, 3100–3105.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Matthews, D. A., Dragovich, P. S., Webber, S. E., Fuhrman, S. A., Patick, A. K., Zalman, L. S., Hendrickson, T. F., Love, R. A., Prins, T. J., Marakovits, J. T., et al. (1999). *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11000–11007.

Matthews, D. A., Smith, W. W., Ferre, R. A., Condon, B., Budahazi, G., Sisson, W., Villafranca, J. E., Janson, C. A., McElroy, H. E., & Gribskov, C. L. (1994). *Cell*. **77**, 761–771.

McCarthy, D. L., Navarrete, S., Willett, W. S., Babbitt, P. C., & Copley, S. D. (1996). *Biochemistry (Mosc.)*. **35**, 14634–14642.

McCoy, A. J. (2002). *Curr. Opin. Struct. Biol.* **12**, 670–673.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). *J. Appl. Crystallogr.* **40**, 658–674.

McCoy, A. J., Storoni, L. C., & Read, R. J. (2004). *Acta Crystallogr. D Biol. Crystallogr.* **60**, 1220–1228.

McCulloch, K. M., Mukherjee, T., Begley, T. P., & Ealick, S. E. (2010). *Biochemistry (Mosc.)*. **49**, 1226–1235.

McPherson, A., Malkin, A. J., & Kuznetsov, Y. G. (2000). *Annu. Rev. Biophys. Biomol. Struct.* **29**, 361–410.

McPherson, A., Malkin, A., & Kuznetsov, Y. (1995). *Structure*. **3**, 759–768.

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., et al. (2001). *Nature*. **409**, 934–941.

Mead, D. A., Szczesna-Skorupa, E., & Kemper, B. (1986). *Protein Eng.* **1**, 67–74.

Medrano, F. J., Alonso, J., Garcia, J. L., Romero, A., Bode, W., & Gomis-Ruth, F. X. (1997). *EMBO J.* **17**, 1–9.

Metzker, M. L. (2010). *Nat. Rev. Genet.* **11**, 31–46.

Mineev, K. S., Goncharuk, S. A., & Arseniev, A. S. (2014). *FEBS Lett.*

Minton, N. P. (1984). *Gene*. **31**, 269–273.

Miyaguchi, K. (2014). *Biol. Cell Auspices Eur. Cell Biol. Organ.* **106**, 323–345.

Modlin, R. L. & Bloom, B. R. (2013). *Sci. Transl. Med.* **5**, 213sr6.

Möller, S., Croning, M. D., & Apweiler, R. (2001). *Bioinforma. Oxf. Engl.* **17**, 646–653.

Morita, Y., Tomida, J., & Kawamura, Y. (2015). *Front. Microbiol.* **6**, 8.

Mueller, J. W., Link, N. M., Matena, A., Hoppstock, L., Rüppel, A., Bayer, P., & Blankenfeldt, W. (2011). *J. Am. Chem. Soc.* **133**, 20096–20099.

Mueller, U., Darowski, N., Fuchs, M. R., Forster, R., Hellmig, M., Paithankar, K. S., Puhringer, S., Steffien, M., Zocher, G., & Weiss, M. S. (2012). *J. Synchrotron Radiat.* **19**, 442–449.

Mukherjee, S. & Hooper, L. V. (2015). *Immunity*. **42**, 28–39.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1992). *Biotechnol. Read. Mass*. **24**, 17–27.

Murayama, T. & Kobayashi, T. (2014). *Methods Mol. Biol. Clifton NJ*. **1177**, 151–161.

Nardini, M., Lang, D. A., Liebeton, K., Jaeger, K.-E., & Dijkstra, B. W. (2000). *J. Biol. Chem.* **275**, 31219–31225.

Nathaniel, C., Wallace, L. A., Burke, J., & Dirr, H. W. (2003). *Biochem. J.* **372**, 241–246.

Nobbmann, U., Connah, M., Fish, B., Varley, P., Gee, C., Mulot, S., Chen, J., Zhou, L., Lu, Y., Shen, F., et al. (2007). *Biotechnol. Genet. Eng. Rev.* **24**, 117–128.

Nurizzo, D., Mairs, T., Guijarro, M., Rey, V., Meyer, J., Fajardo, P., Chavanne, J., Biasci, J. C., McSweeney, S., & Mitchell, E. (2006a). *J. Synchrotron Radiat.* **13**, 227–238.

Nurizzo, D., Mairs, T., Guijarro, M., Rey, V., Meyer, J., Fajardo, P., Chavanne, J., Biasci, J.-C., McSweeney, S., & Mitchell, E. (2006b). The ID23-1 structural biology beamline at the ESRF.

Oakley, A. J. (2005). *Curr. Opin. Struct. Biol.* **15**, 716–723.

Ochsner, U. A., Snyder, A., Vasil, A. I., & Vasil, M. L. (2002). *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8312–8317.

O'Donovan, D., Corcoran, G. D., Lucey, B., & Sleator, R. D. (2014). *Virulence*. **5**, 498–506.

Okai, M., Miyauchi, Y., Ebihara, A., Lee, W. C., Nagata, K., & Tanokura, M. (2007). *Proteins*. **70**, 1646–1649.

Ozer, B., Duran, N., Onlen, Y., & Savas, L. (2012). *J. Antibiot. (Tokyo)*. **65**, 9–13.

Page, W. J. & Doran, J. L. (1981). *J. Bacteriol.* **146**, 33–40.

Panavas, T., Sanders, C., & Butt, T. R. (2009). *Methods Mol. Biol. Clifton NJ*. **497**, 303–317.

Pannu, N. S. & Read, R. J. (2004). *Acta Crystallogr. D Biol. Crystallogr.* **60**, 22–27.

Parsons, J. F., Calabrese, K., Eisenstein, E., & Ladner, J. E. (2003). *Biochemistry (Mosc.)*. **42**, 5684–5693.

Pendleton, J. N., Gorman, S. P., & Gilmore, B. F. (2013). *Expert Rev. Anti Infect. Ther.* **11**, 297–308.

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). *Nat. Methods*. **8**, 785–786.

Peterson, L. R. (2009). *Clin. Infect. Dis.* **49**, 992–993.

Petrella, S., Gelus-Ziental, N., Maudry, A., Laurans, C., Boudjelloul, R., & Sougakoff, W. (2010). *Plos One*. **6**, e15785–e15785.

Phan, J., Zdanov, A., Evdokimov, A. G., Tropea, J. E., Peters, H. K., Kapust, R. B., Li, M., Wlodawer, A., & Waugh, D. S. (2002). *J. Biol. Chem.* **277**, 50564–50572.

Pierce, J. C., Sauer, B., & Sternberg, N. (1992). *Proc. Natl. Acad. Sci. U. S. A.* **89**, 2056–2060.

Porath, J., Carlsson, J., Olsson, I., & Belfrage, G. (1975). *Nature*. **258**, 598–599.

Poteete, A. R., Rosadini, C., & St Pierre, C. (2006). *BioTechniques*. **41**, 261–262, 264.

Potterton, E., Briggs, P., Turkenburg, M., & Dodson, E. (2003). *Acta Crystallogr. D Biol. Crystallogr.* **59**, 1131–1137.

Pruitt, B. A. (2000). *Ann. Surg.* **232**, 287–301.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., et al. (2014). *Nucleic Acids Res.* **42**, D756–D763.

Pruitt, K. D., Brown, G. R., & Tatusova, T. (2002). *The NCBI Handbook*, p. Chapter 18. Bethesda (MD): National Center for Biotechnology Information.

Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). *Nucleic Acids Res.* **40**, D130–D135.

Qiu, D., Damron, F. H., Mima, T., Schweizer, H. P., & Yu, H. D. (2008). *Appl. Environ. Microbiol.* **74**, 7422–7426.

Rajendran, C., Dworkowski, F. S. N., Wang, M., & Schulze-Briese, C. (2011). *J. Synchrotron Radiat.* **18**, 318–328.

Reinemer, P., Dirr, H. W., Ladenstein, R., Schäffer, J., Gallay, O., & Huber, R. (1991). *EMBO J.* **10**, 1997–2005.

Rice, L. B. (2008). *J. Infect. Dis.* **197**, 1079–1081.

Romão, M. J., Turk, D., Gomis-Rüth, F. X., Huber, R., Schumacher, G., Möllering, H., & Rüssmann, L. (1992). *J. Mol. Biol.* **226**, 1111–1130.

Römling, U., Greipel, J., & Tümmler, B. (1995). *Mol. Microbiol.* **17**, 323–332.

Rossjohn, J., McKinstry, W. J., Oakley, A. J., Verger, D., Flanagan, J., Chelvanayagam, G., Tan, K. L., Board, P. G., & Parker, M. W. (1998a). *Struct. Lond. Engl. 1993*. **6**, 309–322.

Rossjohn, J., Polekhina, G., Feil, S. C., Allocati, N., Masulli, M., Di Illio, C., & Parker, M. W. (1998b). *Struct. Lond. Engl. 1993*. **6**, 721–734.

Roy, P. H., Tetu, S. G., Larouche, A., Elbourne, L., Tremblay, S., Ren, Q., Dodson, R., Harkins, D., Shay, R., Watkins, K., et al. (2010). *PloS One*. **5**, e8842.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467.

Saraswat, M., Grand, R. S., & Patrick, W. M. (2013). *Biosci. Biotechnol. Biochem.* **77**, 402–404.

Sawa, T. (2014). *J. Intensive Care*. **2**, 10.

Schirrmeister, B. E., de Vos, J. M., Antonelli, A., & Bagheri, H. C. (2013). *Proc. Natl. Acad. Sci. U. S. A.* **110**, 1791–1796.

Schmidpeter, P. A. M. & Schmid, F. X. (2014). *ACS Chem. Biol.* **9**, 1145–1152.

Schmidt, K. D., Tümmler, B., & Römling, U. (1996). *J. Bacteriol.* **178**, 85–93.

Schmidt, T. G. M., Batz, L., Bonet, L., Carl, U., Holzapfel, G., Kiem, K., Matulewicz, K., Niermeier, D., Schuchardt, I., & Stanar, K. (2013). *Protein Expr. Purif.* **92**, 54–61.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1772–1779.

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). *PLoS Comput. Biol.* **5**,.

Schopf, J. W. (1993). *Science*. **260**, 640–646.

Schopf, J. W. (2006). *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **361**, 869–885.

Schuster, S. C. (2008). *Nat. Methods*. **5**, 16–18.

Sheldrick, G. M. (2010). *Acta Crystallogr. D Biol. Crystallogr.* **66**, 479–485.

Shorr, A. F. (2009). *Crit. Care Med.* **37**, 1463–1469.

Shpyrko, O., Fukuto, M., Pershan, P., Ocko, B., Kuzmenko, I., Gog, T., & Deutsch, M. (2004). *Phys. Rev. B*. **69**, 245423.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). *Mol. Syst. Biol.* **7**, 539.

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. A., Rentzsch, R., et al. (2013). *Nucleic Acids Res.* **41**, D490–D498.

Singh, S. M. & Panda, A. K. (2005). *J. Biosci. Bioeng.* **99**, 303–310.

Sinning, I., Kleywegt, G. J., Cowan, S. W., Reinemer, P., Dirr, H. W., Huber, R., Gilliland, G. L., Armstrong, R. N., Ji, X., & Board, P. G. (1993). *J. Mol. Biol.* **232**, 192–212.

Sivashanmugam, A., Murray, V., Cui, C., Zhang, Y., Wang, J., & Li, Q. (2009). *Protein Sci. Publ. Protein Soc.* **18**, 936–948.

Skopelitou, K., Dhavala, P., Papageorgiou, A. C., & Labrou, N. E. (2012). *PLoS ONE*. **7**,.

Smith, B. C., Anderson, M. A., Hoadley, K. A., Keck, J. L., Cleland, W. W., & Denu, J. M. (2011). *Biochemistry (Mosc.)*. **51**, 243–256.

Smith, B. C., Hallows, W. C., & Denu, J. M. (2009). *Anal. Biochem.* **394**, 101–109.

Smith, D. B. & Johnson, K. S. (1988). *Gene*. **67**, 31–40.

Sørensen, K. I. & Hove-Jensen, B. (1996). *J. Bacteriol.* **178**, 1003–1011.

Sousa, A. M. & Pereira, M. O. (2014). *Pathog. Basel Switz.* **3**, 680–703.

Spencer, D. H., Kas, A., Smith, E. E., Raymond, C. K., Sims, E. H., Hastings, M., Burns, J. L., Kaul, R., & Olson, M. V. (2003). *J. Bacteriol.* **185**, 1316–1325.

Stanberry, L., Rekepalli, B., Liu, Y., Giblock, P., Higdon, R., Montague, E., Broomall, W., Kolker, N., & Kolker, E. (2014). *Concurr. Comput. Pract. Exp.* **26**, 2112–2121.

Stanway, G., Hughes, P. J., Mountford, R. C., Minor, P. D., & Almond, J. W. (1984). *Nucleic Acids Res.* **12**, 7859–7875.

Stein, L. D. (2010). *Genome Biol.* **11**, 207.

Stewart, R. M. K., Wiehlmann, L., Ashelford, K. E., Preston, S. J., Frimmersdorf, E., Campbell, B. J., Neal, T. J., Hall, N., Tuft, S., Kaye, S. B., et al. (2011). *J. Clin. Microbiol.* **49**, 993–1003.

Stourman, N. V., Branch, M. C., Schaab, M. R., Harp, J. M., Ladner, J. E., & Armstrong, R. N. (2011). *Biochemistry (Mosc.)*. **50**, 1274–1281.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., et al. (2000). *Nature*. **406**, 959–964.

Studier, F. W. (1991). *J. Mol. Biol.* **219**, 37–44.

Studier, F. W. (2005). *Protein Expr. Purif.* **41**, 207–234.

Studier, F. W. (2014). *Methods Mol. Biol. Clifton NJ*. **1091**, 17–32.

Studier, F. W. & Moffatt, B. A. (1986). *J. Mol. Biol.* **189**, 113–130.

Sun, Y., Yin, S., Feng, Y., Li, J., Zhou, J., Liu, C., Zhu, G., & Guo, Z. (2014). *J. Biol. Chem.* **289**, 15867–15879.

Sureshan, V., Deshpande, C. N., Boucher, Y., Koenig, J. E., Midwest Center for Structural Genomics, Stokes, H. W., Harrop, S. J., Curmi, P. M. G., & Mabbutt, B. C. (2013). *PloS One*. **8**, e52934.

Svergun, D., Barberato, C., & Koch, M. H. J. (1995). *J. Appl. Crystallogr.* **28**, 768–773.

Tamme, K., Liigant, A., Tapfer, H., & Talvik, R. (2000). *J. Int. Med. Res.* **28**, 199–206.

Tan, H.-L., Wang, Y., Cheng, X.-Q., Huang, Y.-M., Liu, W., & Zhang, L.-J. (2014). *Genome Announc.* **2**, e01280–14.

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., & Tolstoy, I. (2014). *Nucleic Acids Res.* **42**, D553–D559.

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). *Science*. **278**, 631–637.

Taylor, P. L., Blakely, K. M., de Leon, G. P., Walker, J. R., McArthur, F., Evdokimova, E., Zhang, K., Valvano, M. A., Wright, G. D., & Junop, M. S. (2008). *J. Biol. Chem.* **283**, 2835–2845.

Teplyakov, A., Obmolova, G., Badet-Denisot, M. A., Badet, B., & Polikarpov, I. (1998). *Struct. Lond. Engl. 1993*. **6**, 1047–1055.

Terwilliger, T. C. (2000). *Acta Crystallogr. D Biol. Crystallogr.* **56**, 965–972.

Tice, M. M. & Lowe, D. R. (2004). *Nature*. **431**, 549–552.

Tipton, K. A., Coleman, J. P., & Pesci, E. C. (2013). *J. Bacteriol.* **195**, 3433–3441.

Trombe, M. C. (1993). *J. Gen. Microbiol.* **139**, 433–439.

Trombe, M. C., Clavé, C., & Manias, J. M. (1992). *J. Gen. Microbiol.* **138**, 77–84.

Tsuji, A., Kaneko, Y., Takahashi, K., Ogawa, M., & Goto, S. (1982). *Microbiol. Immunol.* **26**, 15–24.

Tullman-Ercek, D., DeLisa, M. P., Kawarasaki, Y., Iranpour, P., Ribnicky, B., Palmer, T., & Georgiou, G. (2007). *J. Biol. Chem.* **282**, 8309–8316.

Tümmler, B., Wiehlmann, L., Klockgether, J., & Cramer, N. (2014). *F1000Prime Rep.* **6**,.

Tu, Z., He, G., Li, K. X., Chen, M. J., Chang, J., Chen, L., Yao, Q., Liu, D. P., Ye, H., Shi, J., et al. (2005). *Electron. J. Biotechnol.* **8**,.

Valadbeigi, H., Tabatabaei, R. R., Malek, A., Sekawi, Z., Raftari, M., Parvaneh, K., Ghafourian, S., & Sadeghifard, N. (2014). *Clin. Lab.* **60**, 363–367.

Vanhoof, R., Godard, C., Nulens, E., Nyssen, H. J., Wildemauwe, C., Hubrechts, J. M., Maes, P., & Hannecart-Pokorni, E. (1993). *J. Hosp. Infect.* **24**, 129–138.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). *Science*. **291**, 1304–1351.

Vinothkumar, K. R., Zhu, J., & Hirst, J. (2014). *Nature*.

Vlamis-Gardikas, A., Aslund, F., Spyrou, G., Bergman, T., & Holmgren, A. (1997). *J. Biol. Chem.* **272**, 11236–11243.

Vuilleumier, S., Ivos, N., Dean, M., & Leisinger, T. (2001). *Microbiol. Read. Engl.* **147**, 611–619.

Wade, N. (2010). *N. Y. Times*.

Wadington, M. C., Ladner, J. E., Stourman, N. V., Harp, J. M., & Armstrong, R. N. (2009). *Biochemistry (Mosc.)*. **48**, 6559–6561.

Walker, T. S., Bais, H. P., Déziel, E., Schweizer, H. P., Rahme, L. G., Fall, R., & Vivanco, J. M. (2004). *Plant Physiol.* **134**, 320–331.

Wang, J., Dauter, M., Alkire, R., Joachimiak, A., & Dauter, Z. (2007). *Acta Crystallogr. D Biol. Crystallogr.* **63**, 1254–1268.

Wang, Y., Qiu, L., Ranson, H., Lumjuan, N., Hemingway, J., Setzer, W. N., Meehan, E. J., & Chen, L. (2008). *J. Struct. Biol.* **164**, 228–235.

Weiss, M. S. (2001). *J. Appl. Crystallogr.* **34**, 130–135.

Welsh, M. A., Eibergen, N. R., Moore, J. D., & Blackwell, H. E. (2015). *J. Am. Chem. Soc.* **137**, 1510–1519.

Wheeler, T. J., Clements, J., & Finn, R. D. (2014). *BMC Bioinformatics*. **15**, 7.

Wilce, M. C. & Parker, M. W. (1994). *Biochim. Biophys. Acta*. **1205**, 1–18.

Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., & Hochstrasser, D. F. (1999). *Methods Mol. Biol. Clifton NJ*. **112**, 531–552.

Williams, P. A. & Worsey, M. J. (1976). *J. Bacteriol.* **125**, 818–828.

Willmott, P. R., Meister, D., Leake, S. J., Lange, M., Bergamaschi, A., Böge, M., Calvi, M., Cancellieri, C., Casati, N., Cervellino, A., et al. (2013). *J. Synchrotron Radiat.* **20**, 667–682.

Winsor, G. L., Lam, D. K. W., Fleming, L., Lo, R., Whiteside, M. D., Yu, N. Y., Hancock, R. E. W., & Brinkman, F. S. L. (2011). *Nucleic Acids Res.* **39**, D596–D600.

Winsor, G. L., Lo, R., Ho Sui, S. J., Ung, K. S. E., Huang, S., Cheng, D., Ching, W.-K. H., Hancock, R. E. W., & Brinkman, F. S. L. (2005). *Nucleic Acids Res.* **33**, D338–D343.

Wolf, Y. I., Makarova, K. S., Yutin, N., & Koonin, E. V. (2012). *Biol. Direct*. **7**, 46.

Woodcock, D. M., Crowther, P. J., Doherty, J., Jefferson, S., DeCruz, E., Noyer-Weidner, M., Smith, S. S., Michael, M. Z., & Graham, M. W. (1989). *Nucleic Acids Res.* **17**, 3469–3478.

Wyss, M. & Kaddurah-Daouk, R. (2000). *Physiol. Rev.* **80**, 1107–1213.

Xiao, G., Liu, S., Ji, X., Johnson, W. W., Chen, J., Parsons, J. F., Stevens, W. J., Gilliland, G. L., & Armstrong, R. N. (1996). *Biochemistry (Mosc.)*. **35**, 4753–4765.

Xing, Y., Li, Z., Chen, Y., Stock, J. B., Jeffrey, P. D., & Shi, Y. (2008). *Cell*. **133**, 154–163.

Xu, N., Yu, S., Moniot, S., Weyand, M., & Blankenfeldt, W. (2012). *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **68**, 1034–1039.

Yahr, T. L., Vallis, A. J., Hancock, M. K., Barbieri, J. T., & Frank, D. W. (1998). *Proc. Natl. Acad. Sci. U. S. A.* **95**, 13899–13904.

Ymer, S. (1991). *Nucleic Acids Res.* **19**, 6960.

Young, C. L., Britton, Z. T., & Robinson, A. S. (2012). *Biotechnol. J.* **7**, 620–634.

Yuan, S. J. & Pehkonen, S. O. (2007). *Colloids Surf. B Biointerfaces*. **59**, 87–99.

Zheng, X., Guo, J., Xu, L., Li, H., Zhang, D., Zhang, K., Sun, F., Wen, T., Liu, S., & Pang, H. (2011). *PloS One*. **6**, e20506.

Zwart, P. H., Afonine, P. V., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., McKee, E., Moriarty, N. W., Read, R. J., Sacchettini, J. C., et al. (2008). *Methods Mol. Biol. Clifton NJ*. **426**, 419–435.

(2010). *Nature*. **464**, 649–650.